R Project – Total Salary Compensation Prediction

Team:
Shephali Jain
Teja Alluru
Shilpi Karmakar
Marion LaRoque
Shalini Bagadhi

INDEX OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

1 ABSTRACT

The main purpose/objective, the methodology used, the results, and the conclusion for executing the project are listed in the sections below.

1.1 Purpose

Emphasizing the higher salaries associated with attracting new analytics professionals, it becomes necessary to understand the patterns in the total salary compensation offered by various companies. With record-breaking attrition rates that the world is seeing as a direct result of the pandemic, it gives a huge bargaining power for graduates in the total salary compensation negotiations. Our team's purpose is to create awareness through ML programs in identifying the total compensation that the market can offer to the graduate students to a high degree of accuracy.

1.2 Benefit

Having the ability to accurately predict the total salary compensation helps in the following ways-

1. To gain a better understanding of what variables affect the compensation and how they can improve upon those variables.
2. To identify what companies, offer better prospects in terms of total salary compensation.

1.3 Methodology

1.3.1 Business/Project Methodology

To execute the project, we will be following the CRISP-DM methodology of AGILE project management. The CRISP-DM methodology used in executing the project is shown in Figure 1-1 and the various stages of the project are shown in Figure 1-2.

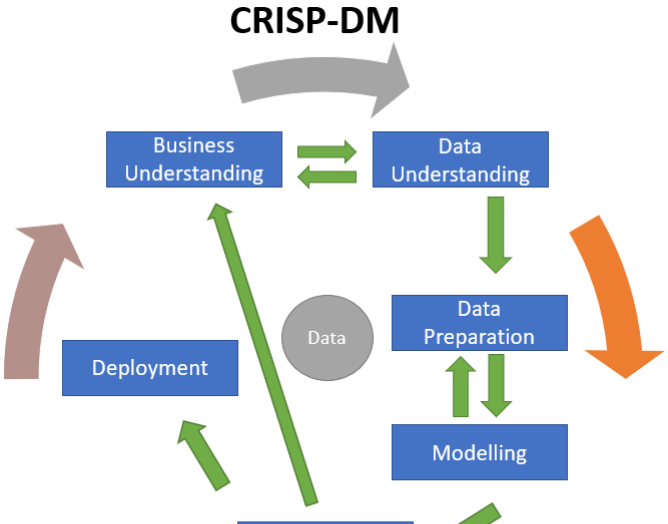


Figure 1-1 Data Analytics Project CRISP-DM Methodology Ref. [3]

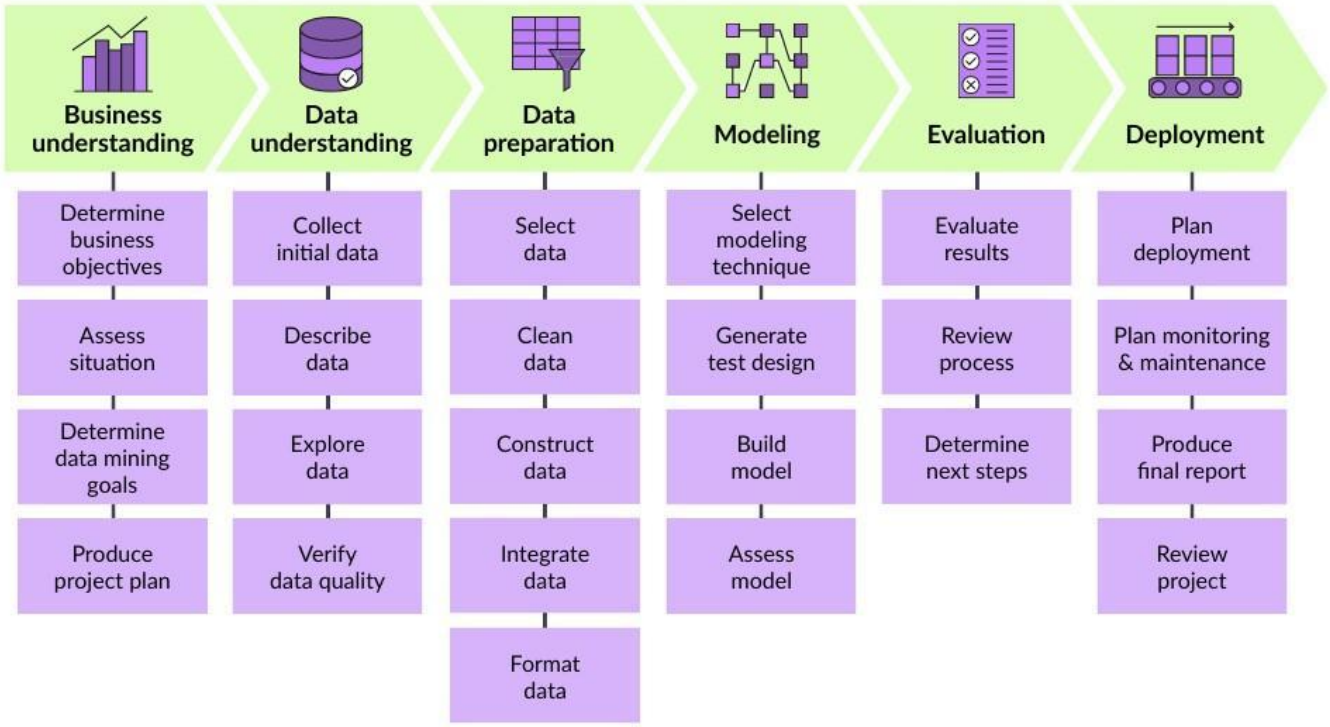


Figure 1-2 CRISP-DM – Stages Ref. [4]

1.3.2 Data Analysis Method

The data analysis methods that we will use are as follows:

1. Plot Scatterplots, Histograms, and Residual Plots – these will check for any non-linearity in the data
2. Categorical Variable Cleaning – performing any dummy coding or collapsing any categories
3. Data Transformation – if any non-linearity is found (may use logarithmic transformation etc.)
4. Perform Ordinary Least Squares (OLS) Regression
5. Check for significant interaction and collinearity
6. Perform post regression analysis of model and individual variable residuals (heteroskedasticity, Cook's distance, etc.)
7. Analyze explanatory variables significance using p-values and use VIF or partial F tests to determine if a variable is significant and could be excluded from the model
8. Determine parsimonious model

1.3.3 Software/ Programs

We will utilize .csv files for the original data source. Statistical/Data Analysis software R is used for data understanding, description, model analysis, and evaluating the model. Microsoft Word and Excel are used for the project documentation part.

1.4 Conclusion/Recommendation

We recommend the graduate students to apply to the top 25 companies which pay higher salaries than others. Also, to check the years of experience to understand what salary they can negotiate for. By going through this report and running their statistics in R they should be able to predict what salary they can negotiate for to a ballpark figure.

2 INTRODUCTION / BUSINESS UNDERSTANDING

2.1 PROJECT DETAILS

This section gives details about the Project i.e., Project background understanding and the problems, challenges, and opportunities that must be addressed by the project and the scope of the project.

2.1.1 Brief Project Background/Understanding

One of the most important aspects of any job search is to know what our experience and education, are worth in the market. A job seeker needs to know what the prevailing trends in the market and how different variables like age, experience, gender, education, and race are affect the total compensation.
Understanding the patterns behind the salary will equip the students with better bargaining power in job offers. This project is to help the students to predict what salaries they can be offered in the market and what they can bargain for.

2.1.2 Research Question/Project Scope

Total Yearly Salary Prediction:
To identify patterns from salary data and to create a machine learning model to predict the total salary compensation for the students.

2.2 PROJECT GOAL AND OBJECTIVES

2.2.1 Goal:

To create a machine learning model to predict total salary compensation.

2.2.2 Objectives:

To determine at least 5-6 variables responsible for indicating total salary compensation of employees using predictive analytics, by analyzing and digging deep-dive in historical salary data. To use any trends that they show as it links to a key in predicting what different employers might offer in the future. The key metrics identification will be quoted for further improvement tactics to be defined for a better model reliable for our goal.

3 DATA UNDERSTANDING/DATA PREPARATION

3.1 Initial Data Collection Report

Efforts has been in continuous research to collect salary data from the global internet. Having worked through several days, team was finally able to get the data for this project from Kaggle .com Ref.[1] and levels.fyi Ref. [2]. The data consists of location, demographics, education, company, and various other data. The dataset also includes the salary details of various companies across the globe.

The complete dataset consists of one .csv file “Levels_Fyi_Salary_Data.csv”. There are no problems recorded while obtaining the data. A more in-depth understanding of the data is established in the chapters below.

3.2 Data Description

The data is contained in one flat file and consists of mainly 62642 records. The number of features, records, and the detailed description of the features in the .csv file listed in section 2.1 is summarized in the data dictionary presented in section 2.2.1 below.

3.2.1 Data Dictionary

The data dictionary for the salary data from “Levels_Fyi_Salary_Data.xlsx” is presented in Table 3-1 below. The initial dataset consists of 29 variables and 62642 observations.

Feature	Description	Data Type	No of Records
timestamp	Date and Time the data was collected	TimeStamp	62642
company	Company for which the data is reported	Text/VarChar	62642
level	Designation/ level	Text/VarChar	62642
title	Job title	Text/VarChar	62642
totalyearlycompensation	Total yearly compensation	Numeric/Int	62642
location	Job location	Text/VarChar	62642
yearsofexperience	Job holder total experience	Numeric/Float	62642
yearsatcompany	Job holder experience in the company	Numeric/Float	62642
tag	Miscellaneous random data	Text/VarChar	62642
basesalary	The base salary for the job	Numeric/Int	62642
stockgrantvalue	Stock value granted for the job	Numeric/Int	62642
bonus	Bonus offered for the job	Numeric/Int	62642
gender	Gender of the job holder	Text/VarChar	62642
otherdetails	Miscellaneous random data	Text/VarChar	62642
cityid	City identification number	Numeric/Int	62642
dmaid	State identification number	Numeric/Int	62642
rowNumber	The row number of instances	Numeric/Int	62642
Masters_Degree	Education of job holder: Masters 1=Yes, 0 = No	Numeric/Int	62642
Bachelors_Degree	Education of job holder: Bachelors 1=Yes, 0 = No	Numeric/Int	62642
Doctorate_Degree	Education of job holder: Doctorate 1=Yes, 0 = No	Numeric/Int	62642
Highschool	Education of job holder: Highschool 1=Yes, 0 = No	Numeric/Int	62642
Some_College	Education of job holder: Some_College 1=Yes, 0 = No	Numeric/Int	62642
Race_Asian	Race of the job holder: Asian 1=Yes, 0 = No	Numeric/Int	62642
Race_White	Race of the job holder: White 1=Yes, 0 = No	Numeric/Int	62642
Race_Two_Or_More	Race of the job holder: Two or More 1=Yes, 0 = No	Numeric/Int	62642
Race_Black	Race of the job holder: Black 1=Yes, 0 = No	Numeric/Int	62642
Race_Hispanic	Race of the job holder: Hispanic 1=Yes, 0 = No	Numeric/Int	62642
Race	Race of the job holder	Text/VarChar	62642
Education	Education of job holder	Text/VarChar	62642

Table 3-1: Data Dictionary: Salary Data

3.3 Data Exploration

The initial data exploration was carried out in R using summary statistics & visualizations and the summary statistics & univariate properties are listed in section 3.3.2 and section 3.3.3 respectively.

The summary descriptive statistics of the raw data are presented in Figure 3-1 below.

```
  timestamp                     company            level              title          totalyearlycompensation
Min.   :2017-06-07 11:33:27   Length:62642       Length:62642       Length:62642       Min.   :  10000
1st Qu.:2020-01-11 23:12:55   Class :character   Class :character   Class :character   1st Qu.: 135000
Median :2020-09-21 16:31:21   Mode  :character   Mode  :character   Mode  :character   Median : 188000
Mean   :2020-07-17 03:30:08                                                           Mean   : 216300
3rd Qu.:2021-03-26 15:19:06                                                           3rd Qu.: 264000
Max.   :2021-08-17 08:28:57                                                           Max.   :4980000

  location         yearsofexperience yearsatcompany     tag             basesalary       stockgrantvalue
Length:62642       Min.   : 0        Min.   : 0     Length:62642       Min.   :     0   Min.   :      0
Class :character   1st Qu.: 3        1st Qu.: 0     Class :character   1st Qu.: 108000   1st Qu.:      0
Mode  :character   Median : 6        Median : 2     Mode  :character   Median : 140000   Median :  25000
                   Mean   : 7        Mean   : 3                        Mean   : 136687   Mean   :  51486
                   3rd Qu.:10        3rd Qu.: 4                        3rd Qu.: 170000   3rd Qu.:  65000
                   Max.   :69        Max.   :69                        Max.   :1659870   Max.   :2800000

    bonus            gender          otherdetails          cityid          dmaid       rowNumber
Min.   :      0   Length:62642       Length:62642       Min.   :    0   Min.   :  0   Min.   :    1
1st Qu.:   1000   Class :character   Class :character   1st Qu.: 7369   1st Qu.:506   1st Qu.:20069
Median :  14000   Mode  :character   Mode  :character   Median : 7839   Median :807   Median :42019
Mean   :  19335                                         Mean   : 9856   Mean   :616   Mean   :41695
3rd Qu.:  26000                                         3rd Qu.:11521   3rd Qu.:807   3rd Qu.:63022
Max.   :1000000                                         Max.   :47926   Max.   :881   Max.   :83875
                                                                        NA's   :2
Masters_Degree Bachelors_Degree Doctorate_Degree   Highschool    Some_College    Race_Asian     Race_White
Min.   :0.00   Min.   :0.0      Min.   :0.00     Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.00
1st Qu.:0.00   1st Qu.:0.0      1st Qu.:0.00     1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00
Median :0.00   Median :0.0      Median :0.00     Median :0.00   Median :0.00   Median :0.00   Median :0.00
Mean   :0.25   Mean   :0.2      Mean   :0.03     Mean   :0.01   Mean   :0.01   Mean   :0.19   Mean   :0.13
3rd Qu.:0.00   3rd Qu.:0.0      3rd Qu.:0.00     3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00
Max.   :1.00   Max.   :1.0      Max.   :1.00     Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.00

Race_Two_Or_More   Race_Black   Race_Hispanic      Race           Education
Min.   :0.00     Min.   :0.00   Min.   :0.00   Length:62642       Length:62642
1st Qu.:0.00     1st Qu.:0.00   1st Qu.:0.00   Class :character   Class :character
Median :0.00     Median :0.00   Median :0.00   Mode  :character   Mode  :character
Mean   :0.01     Mean   :0.01   Mean   :0.02
3rd Qu.:0.00     3rd Qu.:0.00   3rd Qu.:0.00
Max.   :1.00     Max.   :1.00   Max.   :1.00
```

Figure 3-1: Descriptive Statistics – Raw Data

3.3.1 Initial Trimming of Data

Based on the initial visual data exploration from the .csv file and the data dictionary provided in Table 2-1 the following observations were found.

The level data is dropped outright as each company has a different level designated to their organizations and it mostly depends on the experience. So, dropping the column would not take out relevance/significance of levels in the prediction algorithm as it will be captured based on the total experience of the job holder.

The location data is further split into three variables in total consisting of city, state, and country variables. The cityid and dmaid are also removed as they are the numeric identity numbers of the city and state of the position. The city and state details are already recorded.

Row numbers do not make any sense as such are removed.

Education variables consist of the required data and the data is repeated in five more columns as each of the categories in the education variable is again represented in a

different column with 0's and 1's. Anyhow, we will be creating dummy variables during our analysis right now the other five columns are dropped and only the education variable is retained.

Following the similar methodology described in the above point, for the race, only one column is retained, and the other five columns are dropped.

All the initial data trimming is carried out in excel and no software is used, as for this project it seems cost & time effective

3.3.2 Summary Statistics – Trimmed Data

The summary statistics of the trimmed data are presented in Figure 3-2 below. As you can observe from Figure 3-2 the NAs are not recognized as nulls and are part of the Class category as such R code is run and the NAs are presented in Figure 3-3.

```
  timestamp                     company            title          totalyearlycompensation     city
Min.   :2017-06-07 11:33:27   Length:62642       Length:62642       Min.   :  10000         Length:62642
1st Qu.:2020-01-11 23:12:55   Class :character   Class :character   1st Qu.: 135000         Class :character
Median :2020-09-21 16:31:21   Mode  :character   Mode  :character   Median : 188000         Mode  :character
Mean   :2020-07-17 03:30:08                                        Mean   : 216300
3rd Qu.:2021-03-26 15:19:06                                        3rd Qu.: 264000
Max.   :2021-08-17 08:28:57                                        Max.   :4980000
   state              country          yearsofexperience yearsatcompany     tag              basesalary
Length:62642       Length:62642       Min.   : 0        Min.   : 0     Length:62642       Min.   :     0
Class :character   Class :character   1st Qu.: 3        1st Qu.: 0     Class :character   1st Qu.: 108000
Mode  :character   Mode  :character   Median : 6        Median : 2     Mode  :character   Median : 140000
                                      Mean   : 7        Mean   : 3                        Mean   : 136687
                                      3rd Qu.:10        3rd Qu.: 4                        3rd Qu.: 170000
                                      Max.   :69        Max.   :69                        Max.   :1659870
stockgrantvalue       bonus            gender             Race           Education
Min.   :      0   Min.   :      0   Length:62642       Length:62642       Length:62642
1st Qu.:      0   1st Qu.:   1000   Class :character   Class :character   Class :character
Median :  25000   Median :  14000   Mode  :character   Mode  :character   Mode  :character
Mean   :  51486   Mean   :  19335
3rd Qu.:  65000   3rd Qu.:  26000
Max.   :2800000   Max.   :1000000
            timestamp                company                  title totalyearlycompensation
                    0                      0                      0                      0
                 city                  state                country       yearsofexperience
                    0                      0                      0                      0
       yearsatcompany                    tag             basesalary         stockgrantvalue
                    0                      0                      0                      0
                bonus                 gender                   Race              Education
                    0                      0                      0                      0
```

Figure 3-2: Descriptive Statistics – Trimmed Data

```
SalaryData2 <- data.frame(SalaryData)
make.true.NA <- function(x) if(is.character(x)||is.factor(x)){
                                is.na(x) <- x=="NA"; x} else {
                                x}
SalaryData2[] <- lapply(SalaryData2, make.true.NA)

summary(SalaryData2)

sapply(SalaryData2, function(x) sum(is.na(x)))
```

```
     timestamp                company                  title totalyearlycompensation 
             0                      0                      0                      0 
          city                  state                country       yearsofexperience 
             0                      5                      0                      0 
yearsatcompany                    tag             basesalary          stockgrantvalue 
             0                    808                      0                      0 
         bonus                 gender                   Race              Education 
             0                  19540                  40215                  32272 
```

Figure 3-3: NA Values – Trimmed Data

Since we have a vast number of records of 62642, we will go ahead and remove the NA rows to prepare our data for modeling. The R code is used to omit the NA values and the revised summary statistics of the new dataset are presented in Figure 3-4

```
   timestamp                     company             title          totalyearlycompensation     city          
 Min.   :2020-01-27 22:59:06   Length:21580       Length:21580       Min.   :  10000         Length:21580      
 1st Qu.:2020-10-29 22:23:08   Class :character   Class :character   1st Qu.: 119000         Class :character  
 Median :2021-03-03 02:03:06   Mode  :character   Mode  :character   Median : 174000         Mode  :character  
 Mean   :2021-02-15 08:02:13                                        Mean   : 197854                           
 3rd Qu.:2021-05-29 03:55:53                                        3rd Qu.: 245000                           
 Max.   :2021-08-17 08:28:57                                        Max.   :4980000                           
    state              country          yearsofexperience yearsatcompany      tag              basesalary      
 Length:21580       Length:21580       Min.   : 0        Min.   : 0       Length:21580       Min.   :  4000  
 Class :character   Class :character   1st Qu.: 3        1st Qu.: 0       Class :character   1st Qu.:100000  
 Mode  :character   Mode  :character   Median : 6        Median : 2       Mode  :character   Median :135000  
                                       Mean   : 7        Mean   : 3                           Mean   :133872  
                                       3rd Qu.:10        3rd Qu.: 4                           3rd Qu.:165000  
                                       Max.   :45        Max.   :40                           Max.   :900000  
 stockgrantvalue       bonus           gender             Race              Education        
 Min.   :     0   Min.   :     0   Length:21580       Length:21580       Length:21580      
 1st Qu.:     0   1st Qu.:  3000   Class :character   Class :character   Class :character  
 Median : 20000   Median : 13000   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 44885   Mean   : 18418                                                          
 3rd Qu.: 55000   3rd Qu.: 25000                                                          
 Max.   :954000   Max.   :900000                                                          
```

```
            timestamp                 company                   title totalyearlycompensation 
                    0                       0                       0                       0 
                 city                   state                 country       yearsofexperience 
                    0                       0                       0                       0 
       yearsatcompany                     tag              basesalary          stockgrantvalue 
                    0                       0                       0                       0 
                bonus                  gender                    Race               Education 
                    0                       0                       0                       0 
```

Figure 3-4: Salary Dataset – Null Values Removed

3.4 Condensing/Filtering Categorical Variables

Having observed the initial statistics for the data after the null values were removed. Based on the statistics shown in Figure 3-4, we have decided to condense some of the categorical variables and filter some of the variables in the dataset. This is carried out to make sure that our regression model runs smoothly and efficiently.

3.4.1 Location Data

Since the data we are interested in is only about salary details of companies located in the USA, we have filtered the data for country == USA. Further, the location data is condensed into nine categorical regions based on the geographic wealth of the different regions in the USA as shown in Figure 3-5. The categorization of regions based on wealth is taken from Ref.[5].

```
# Creating 9 regions - Based on geographic wealth
NE.ref <- c("CT", "MA", "ME", "NH", "RI", "VT")
MA.ref <- c("NJ", "NY", "PA")
ENC.ref <- c("WI", "IL", "IN", "MI", "OH")
WNC.ref <- c("IA", "MN", "MO","KS", "ND", "NE", "SD")
SA.ref <- c("DE", "MD", "WV", "VA", "NC", "SC", "GA", "FL", "DC")
ESC.ref <- c("KY", "TN", "AL", "MS")
WSC.ref <- c("AR","LA", "OK", "TX")
M.ref <- c("MT", "ID", "WY", "NV", "UT", "CO", "AZ", "NM")
P.ref <- c("AK", "CA", "HI", "OR", "WA")

NineRegion.list <- list(
  NorthEast = NE.ref,
  MidAtlantic = MA.ref,
  EastNorthCentral = ENC.ref,
  WestNorthCentral = WNC.ref,
  SouthAtlantic = SA.ref,
  EastSouthCentral = ESC.ref,
  WestSouthCentral = WSC.ref,
  Mountain = M.ref,
  Pacific = P.ref
)

SalaryData6$region <- sapply(SalaryData6$state, function(x) names(NineRegion.list)[grep(x,NineRegion.list)])



SalaryData6$region <- factor(SalaryData6$region)
SalaryRegionCount <- as.data.frame(table(SalaryData6$region))

region <- aggregate(totalyearlycompensation ~ region, SalaryData6, mean)
region[with(region, order(totalyearlycompensation)),]

show(SalaryRegionCount)
```

	region <fctr>	totalyearlycompensation <dbl>
8	WestNorthCentral	129781.0
2	EastSouthCentral	156074.1
1	EastNorthCentral	164056.8
9	WestSouthCentral	168058.7
7	SouthAtlantic	170045.1
4	Mountain	175381.1
5	NorthEast	207969.3
3	MidAtlantic	229005.3
6	Pacific	261560.9

Var1 <fctr>	Freq <int>
EastNorthCentral	264
EastSouthCentral	27
MidAtlantic	1323
Mountain	286
NorthEast	391
Pacific	8429
SouthAtlantic	843
WestNorthCentral	137
WestSouthCentral	835

Figure 3-5: Condensed Location Data – Regions Ref.[5]

3.4.2 Education Data

The education categorical variable has only 4 levels and each of the levels has enough observations as seen in Figure 3-6. Therefore, there is no need to condense the data. However, we have decided to enforce the order based on levels of study.

```
SalaryData6$Education <- factor(SalaryData6$Education)
SalaryEdCount <- as.data.frame(table(SalaryData6$Education))

Education <- aggregate(totalyearlycompensation ~ Education, SalaryData6, mean)
Education[with(Education, order(totalyearlycompensation)),]

show(SalaryEdCount)

#Enforcing order
SalaryData6$Education <- factor(SalaryData6$Education,
                                levels = c("Highschool", "Some College","Bachelor's Degree", "Master's Degree",
"PhD"))
```

	Education <fctr>	totalyearlycompensation <dbl>
1	Bachelor's Degree	219149.2
2	Highschool	244203.0
3	Master's Degree	248512.1
5	Some College	258193.7
4	PhD	313560.0

Var1 <fctr>	Freq <int>
Bachelor's Degree	5863
Highschool	133
Master's Degree	5673
PhD	675
Some College	191

Figure 3-6: Condensed Location Data – Education

3.4.3 Company Data

The company data is filtered based on the frequency, i.e., only companies with observation greater than or equal to 25 are selected as shown in Figure 3-6. This is done as fewer observations don't generate much information related to companies. The total number of companies is now reduced to 96 from 1085. The total number of observations in the dataset after this filter is 12535.

```
count(SalaryData4$company)
SalaryData5<-transform(SalaryData4,Company_Frequency=ave(seq(nrow(SalaryData4)),company,FUN=length))
SalaryData6<- subset(SalaryData5, SalaryData5$Company_Frequency >= 25)
count(SalaryData6$company)
```

Description: df [1,085 x 2]

x <chr>	freq <int>
10x Genomics	2
23andMe	1
2U	5
3M	8
8x8	3
ABB	5
Abbott	5

Description: df [96 x 2]

x <chr>	freq <int>
Accenture	110
Adobe	101
Airbnb	37
Amazon	2077
AMD	47
American Express	39
Apple	563

Figure 3-7: Condensed Location Data – Company

3.4.4 Title Data

The title variable has 15 levels and each of the levels has enough observations as seen in Figure 3-8. Therefore, there is no need to condense the data. However, we have decided to put recruiter at number 1, because we need a business analyst, whichever we put here first will act as a baseline and will not show a coefficient.

```
SalaryData6$title <- factor(SalaryData6$title)
SalarytitleCount <- as.data.frame(table(SalaryData6$title))

title <- aggregate(totalyearlycompensation ~ title, SalaryData6, mean)
title[with(title, order(totalyearlycompensation)),]

show(SalarytitleCount)

SalaryData6$title <- factor(SalaryData6$title,
                            levels = c("Recruiter",
                                       "Business Analyst",
                                       "Mechanical Engineer",
                                       "Management Consultant",
                                       "Human Resources",
                                       "Marketing",
                                       "Data Scientist",
                                       "Sales",
                                       "Hardware Engineer",
                                       "Software Engineer",
                                       "Product Designer",
                                       "Solution Architect",
                                       "Technical Program Manager",
                                       "Product Manager",
                                       "Software Engineering Manager"))
```

	title <fctr>	totalyearlycompensation <dbl>	Var1 <fctr>	Freq <int>
1	Business Analyst	146261.4	Business Analyst	241
10	Recruiter	163805.6	Data Scientist	529
5	Management Consultant	169269.5	Hardware Engineer	507
7	Mechanical Engineer	193639.1	Human Resources	111
4	Human Resources	209486.5	Management Consultant	256
6	Marketing	212398.4	Marketing	246
11	Sales	219791.3	Mechanical Engineer	133
14	Solution Architect	226709.1	Product Designer	401
3	Hardware Engineer	227189.3	Product Manager	971
2	Data Scientist	229589.8	Recruiter	144

Figure 3-8: Condensed Role Data – Title

3.4.5 Gender Data

The gender variable has 3 levels namely male, female and other. The other instance has very few observations as compared to the other two, therefore we have enforced the variable order with others at first so the data for dummy variables created for male and female are retained. The code runs for enforcing the order, the ranking based on total compensation, and the number of observations for each level is presented for reference in Figure 3-9.

```
SalaryData6$gender <- factor(SalaryData6$gender)
SalarygenderCount <- as.data.frame(table(SalaryData6$gender))

gender <- aggregate(totalyearlycompensation ~ gender, SalaryData6, mean)
gender[with(gender, order(totalyearlycompensation)),]

show(SalarygenderCount)

SalaryData6$gender <- factor(SalaryData6$gender,
                             levels = c("Other",
                                        "Female",
                                        "Male"))
```

	gender <fctr>	totalyearlycompensation <dbl>
1	Female	210224.0
2	Male	245356.4
3	Other	289018.9

Var1 <fctr>	Freq <int>
Female	2554
Male	9928
Other	53

Figure 3-9: Condensed Gender Data

3.5 Variable Summary – All

The variable summary for all the variables in consideration is presented in Figure 3-10.

```
  timestamp                     company                                  title
Min.   :2020-07-03 19:56:38   Length:12535       Software Engineer           :7556
1st Qu.:2020-10-21 11:46:13   Class :character   Product Manager             : 971
Median :2021-02-19 11:40:33   Mode  :character   Software Engineering Manager: 624
Mean   :2021-02-06 01:04:50                      Data Scientist              : 529
3rd Qu.:2021-05-18 04:05:11                      Hardware Engineer           : 507
Max.   :2021-08-17 08:28:57                      Technical Program Manager   : 481
                                                 (Other)                     :1867
totalyearlycompensation     city              state             country          yearsofexperience
Min.   :  16000         Length:12535       Length:12535       Length:12535       Min.   : 0.000
1st Qu.: 157000         Class :character   Class :character   Class :character   1st Qu.: 3.000
Median : 205000         Mode  :character   Mode  :character   Mode  :character   Median : 5.000
Mean   : 238383                                                                  Mean   : 7.303
3rd Qu.: 283000                                                                  3rd Qu.:10.000
Max.   :4980000                                                                  Max.   :45.000

yearsatcompany       tag              basesalary     stockgrantvalue      bonus           gender
Min.   : 0.000   Length:12535       Min.   : 13000   Min.   :     0   Min.   :     0   Other :  53
1st Qu.: 0.000   Class :character   1st Qu.:125000   1st Qu.: 10000   1st Qu.:  8000   Female:2554
Median : 2.000   Mode  :character   Median :150000   Median : 35000   Median : 18000   Male  :9928
Mean   : 2.833                      Mean   :153731   Mean   : 61031   Mean   : 22501
3rd Qu.: 4.000                      3rd Qu.:175000   3rd Qu.: 80000   3rd Qu.: 30000
Max.   :40.000                      Max.   :900000   Max.   :954000   Max.   :555000

    Race                      Education      Quarter          Company_Frequency logTotalSalary
Length:12535       Highschool       : 133   Length:12535       Min.   :  25.0    Min.   : 9.68
Class :character   Some College     : 191   Class :character   1st Qu.: 101.0    1st Qu.:11.96
Mode  :character   Bachelor's Degree:5863   Mode  :character   Median : 291.0    Median :12.23
                   Master's Degree  :5673                      Mean   : 728.3    Mean   :12.26
                   PhD              : 675                      3rd Qu.:1290.0    3rd Qu.:12.55
                                                               Max.   :2077.0    Max.   :15.42

            region
Pacific         :8429
MidAtlantic     :1323
SouthAtlantic   : 843
WestSouthCentral: 835
NorthEast       : 391
Mountain        : 286
(Other)         : 428
```

Figure 3-10: Variable Summary – All Variables

3.6 Histograms – Univariate

The univariate histograms/bar plots for all the variables in consideration are presented in Figure 3-11 and Figure 3-12.

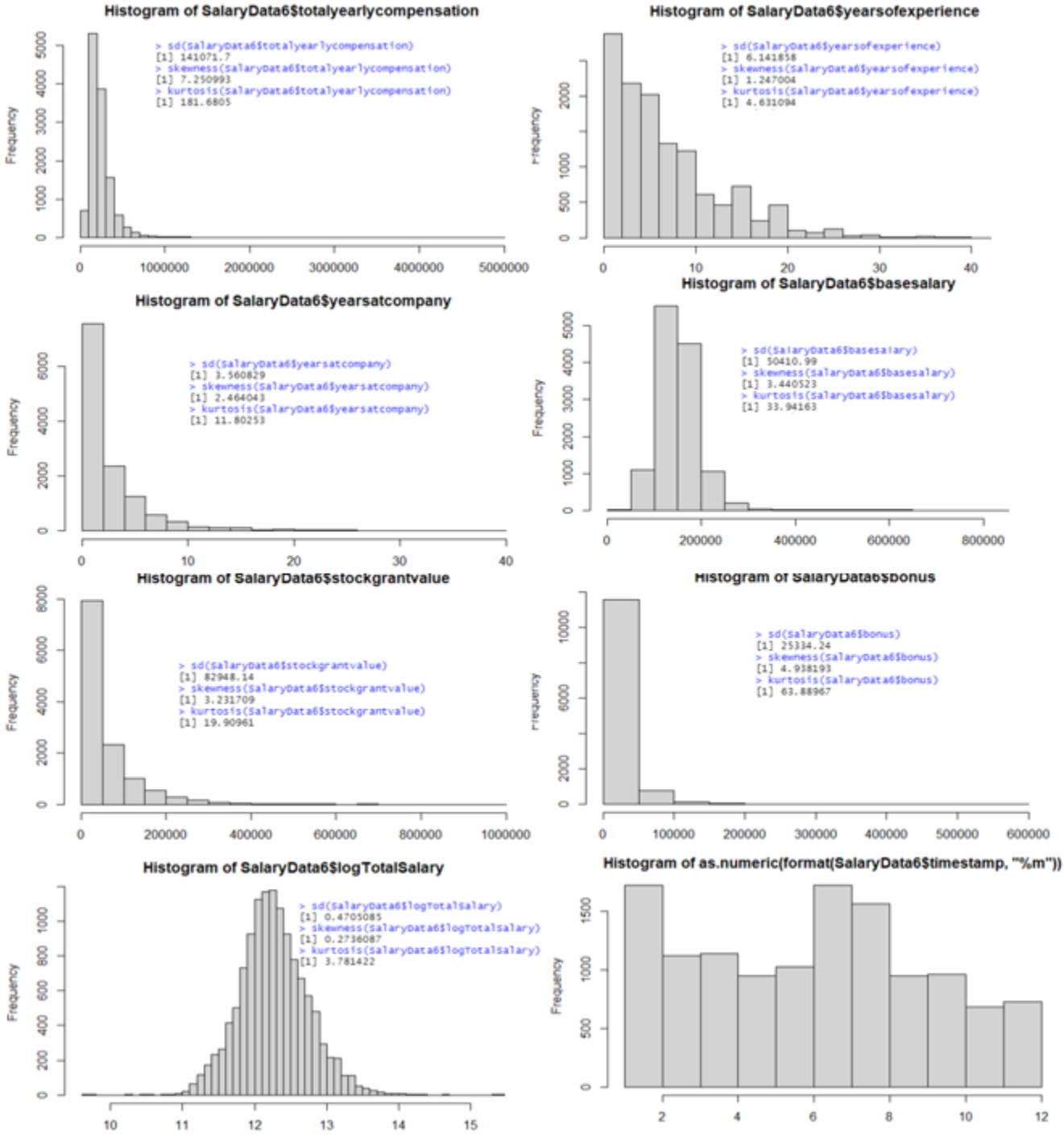


Figure 3-11: Histograms/Bar Plots – Numeric Data

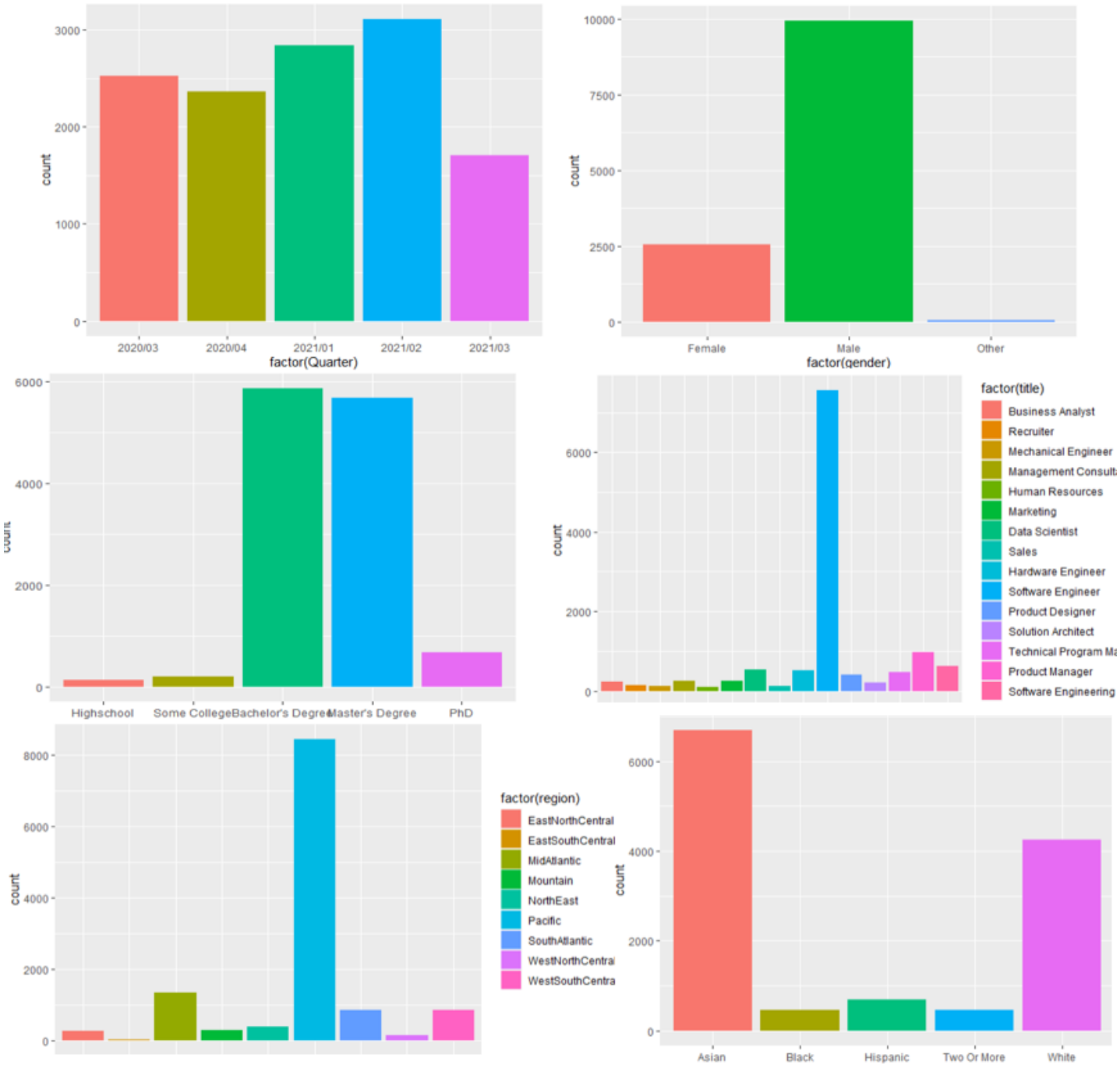


Figure 3-12: Histograms/Bar Plots – Categorical Data

3.7 Descriptive Statistics

The univariate statistics for the features and label for the condensed data are presented in Figure 3-11 below. The visualization and complete summary of the univariate properties are presented in “Univariate Statistics.xlsx”. The total number of observations in the dataset is now 12535.

The basic deductions from Figure 3-11 below are

1. No missing values as all the missing values have been removed from the steps followed in the chapters above.
2. Skewness and kurtosis problem for all variables.

Feature	Type	Data Type	Variable Type	Count	Missing Values	Unique Values
Quarter*	Feature	timestamp	Categorical	12535	0	5
company*	Feature	Text/Char	Categorical	12535	0	96
title	Feature	Text/Char	Categorical	12535	0	15
totalyearlycompensation	Lable	Integer/Double	Continuous	12535	0	661
region*	Feature	Text/Char	Categorical	12535	0	9
yearsofexperience	Feature	Integer/Double	Continuous	12535	0	41
years at company	Feature	Integer/Double	Continuous	12535	0	30
tag	Feature	Text/Char	Categorical	12535	0	1044
basesalary	Feature	Integer/Double	Continuous	12535	0	293
stockgrantvalue	Feature	Integer/Double	Continuous	12535	0	406
bonus	Feature	Integer/Double	Continuous	12535	0	163
gender	Feature	Text/Char	Categorical	12535	0	3
race	Feature	Text/Char	Categorical	12535	0	5
education	Feature	Text/Char	Categorical	12535	0	5

Feature	Mean	Min	Max	Median	Standar Deviation	First Quartile	Third Quartile	Skewness	Excess Kurtosis
Quarter*									
company*									
title									
totalyearlycompensation	238383.00	16000.00	4980000.00	205000.00	141071.70	157000.00	283000.00	7.25	181.68
region*									
yearsofexperience	7.30	0.00	45.00	5.00	6.14	3.00	10.00	1.25	4.63
years at company	2.83	0.00	40.00	2.00	3.56	0.00	4.00	2.46	11.80
tag									
basesalary	153731.00	13000.00	900000.00	150000.00	50410.99	125000.00	175000.00	3.44	33.94
stockgrantvalue	61031.00	0.00	954000.00	35000.00	82948.14	10000.00	80000.00	3.23	19.91
bonus	22501.00	0.00	555000.00	18000.00	25334.24	8000.00	30000.00	4.94	63.89
gender									
race									
education									

* Derived/Filtered Variables

Figure 3-11: Univariate Statistics

3.8 Scatter Plots – Bivariate

The bivariate scatter plots for dependent variable vs independent variables in consideration are presented in Figure 3-14 and Figure 3-15. The bivariate statistics for the log of the dependent variable and independent variable are presented in Figure 3-16 and Figure 3-17.

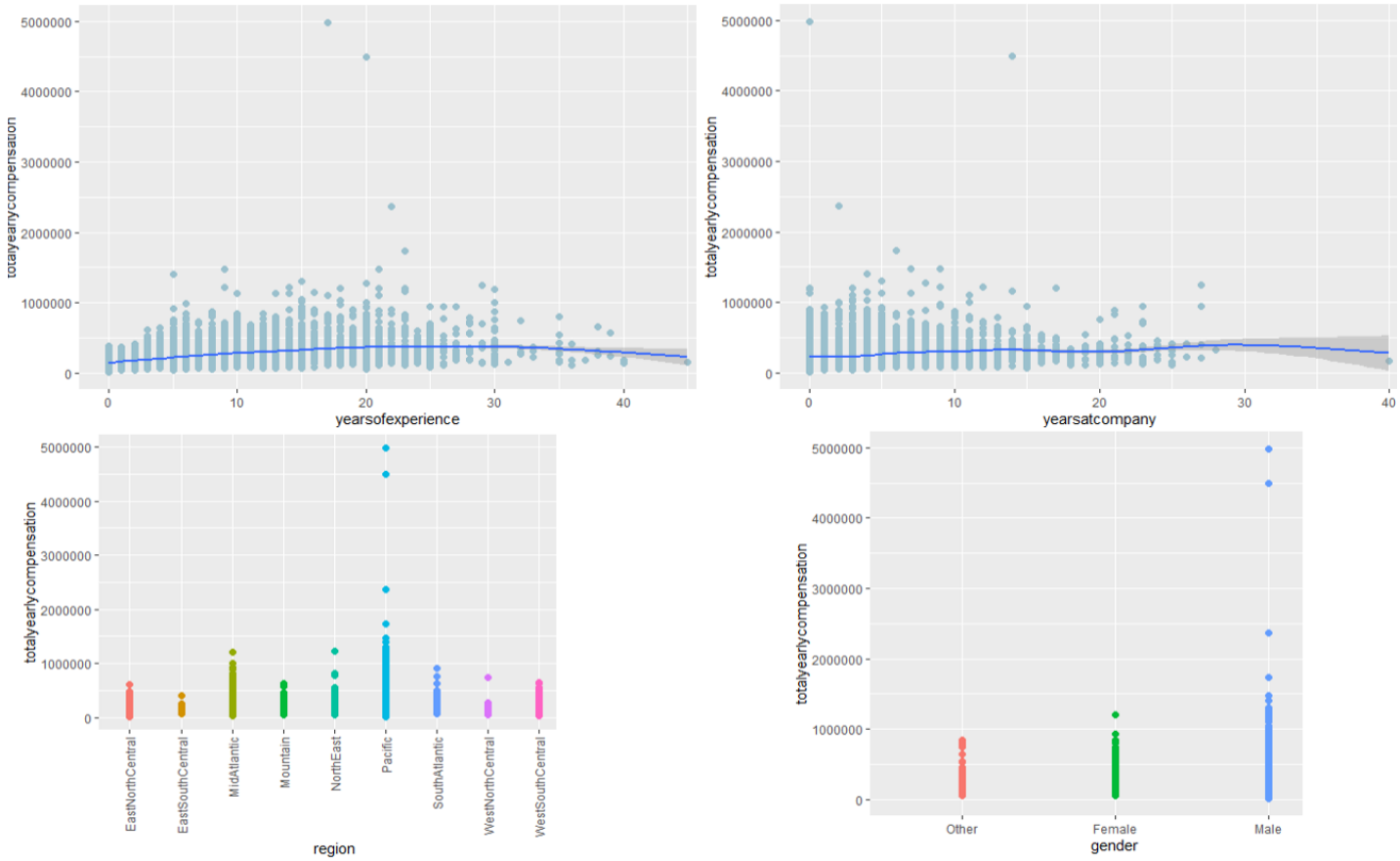


Figure 3-14: Scatter Plots – Dependent Vs Independent

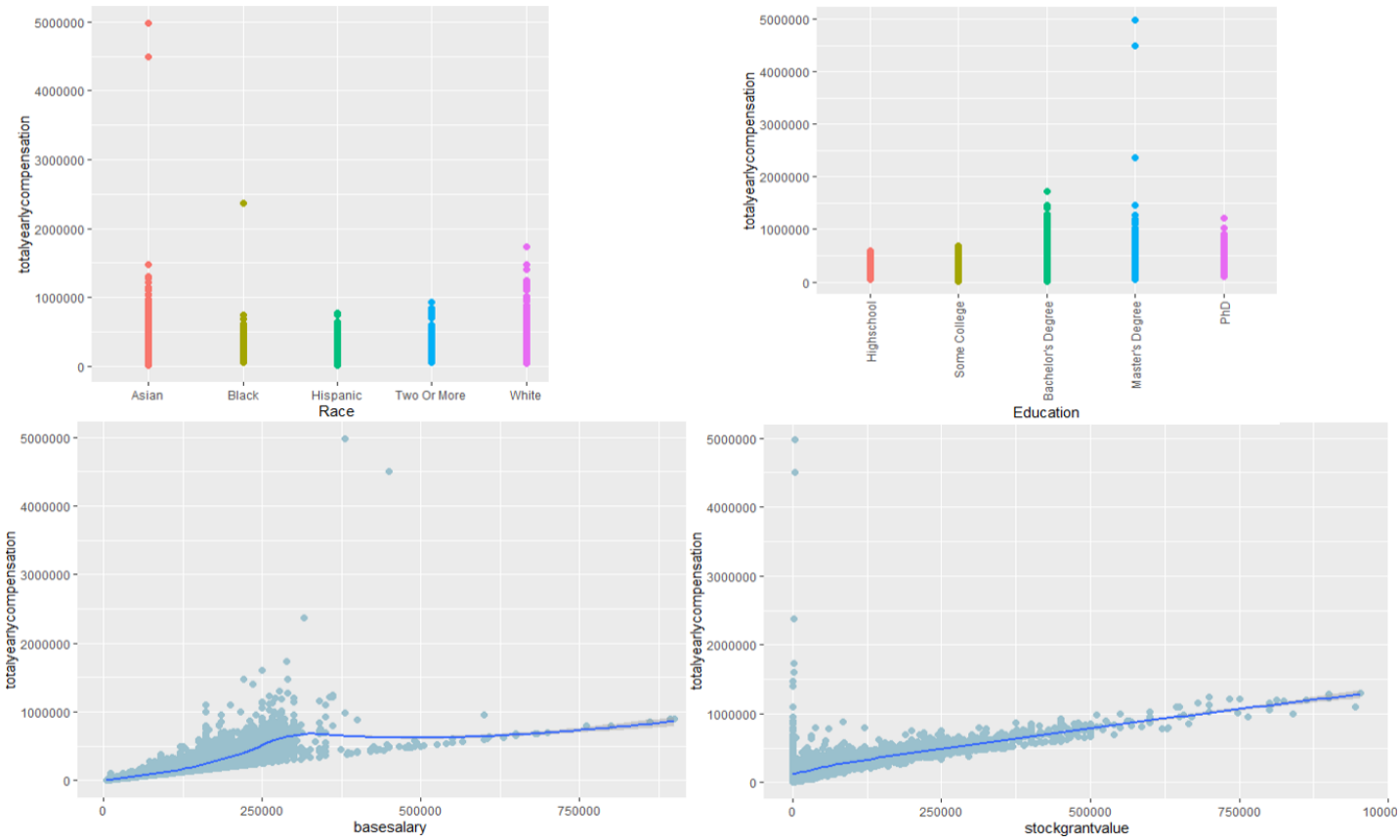


Figure 3-15: Scatter Plots – Dependent Vs Independent

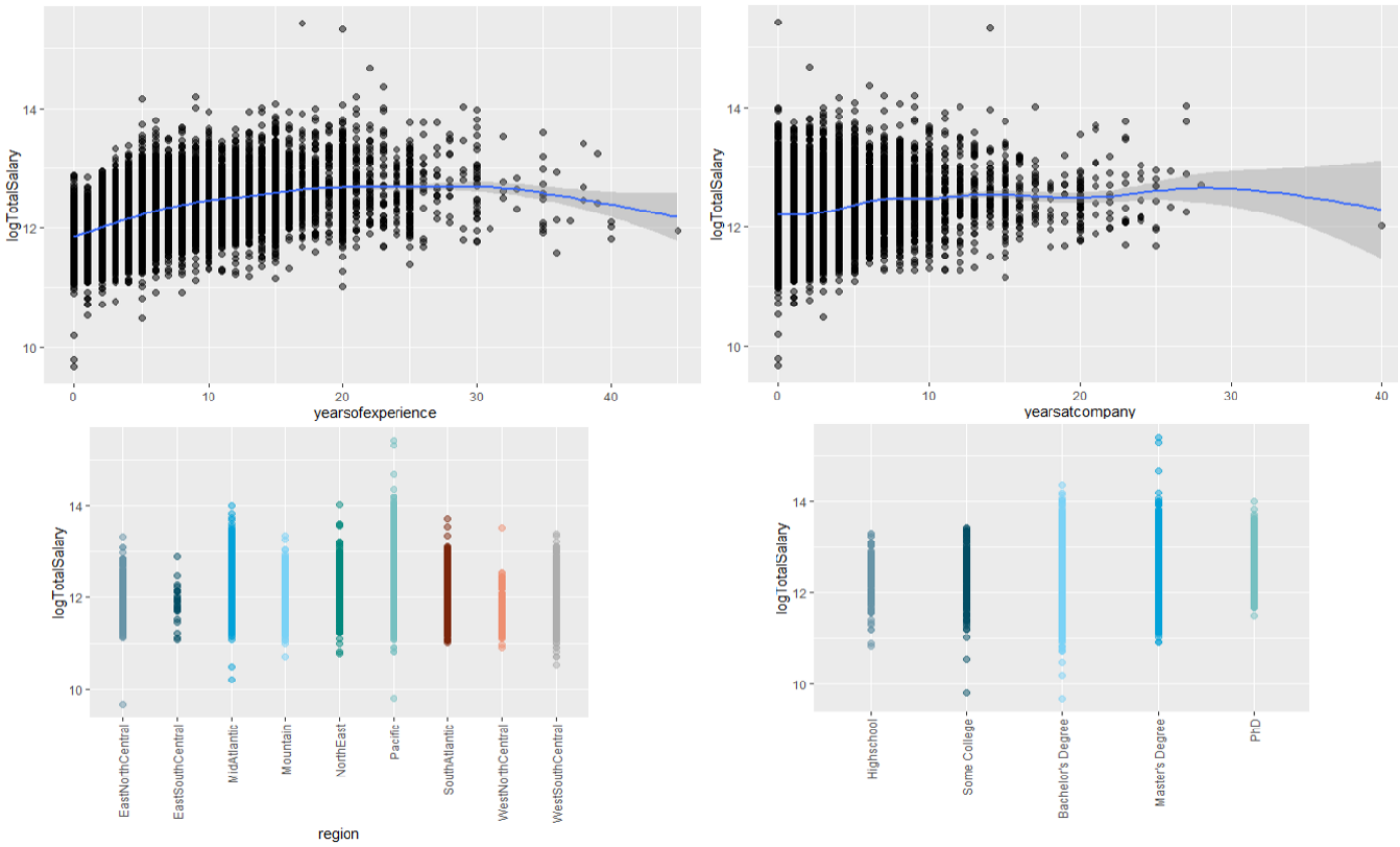


Figure 3-16: Scatter Plots – Dependent Vs Independent – Log Total Salary

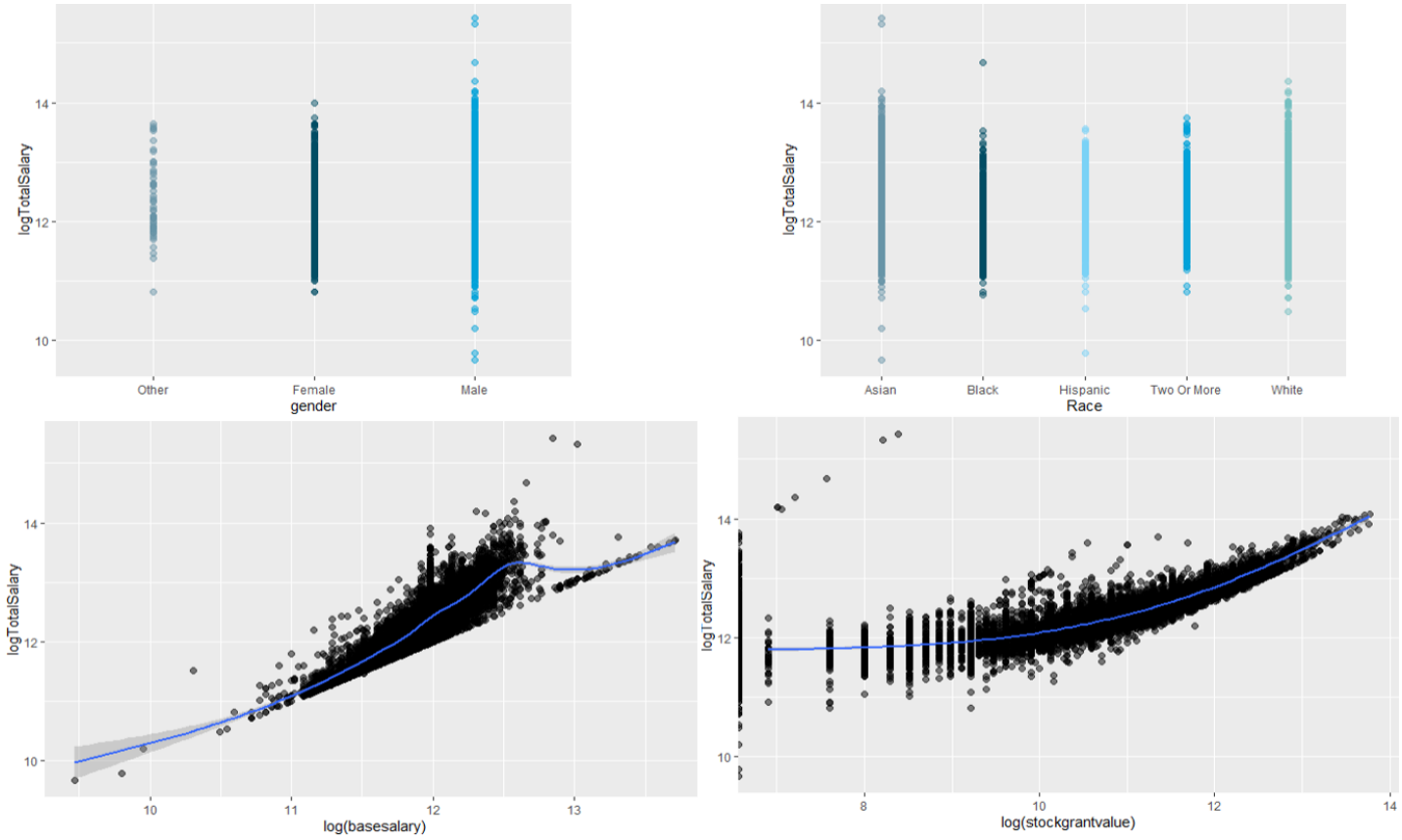


Figure 3-17: Scatter Plots – Dependent Vs Independent - Log Total Salary

3.9 Correlation

Since we are predicting total salary compensation which is a combination of other salary components like base salary, stock grants, and bonus we can ignore the other variables. To cross verify a simple correlation plot and correlation significance table are created in Figure 3-14 and Figure 3-15 respectively. As stated above it's clear from Figure 3-14 that a strong correlation exists between base salary, stock grants, bonus, and totalyearly compensation as total yearly compensation is derived by adding the others.

```
numericdata <- SalaryData6 %>% select(totalyearlycompensation , basesalary , stockgrantvalue , bonus,
yearsofexperience,yearsatcompany )
library(corrplot)
library(RColorBrewer)
M <-cor(numericdata)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```

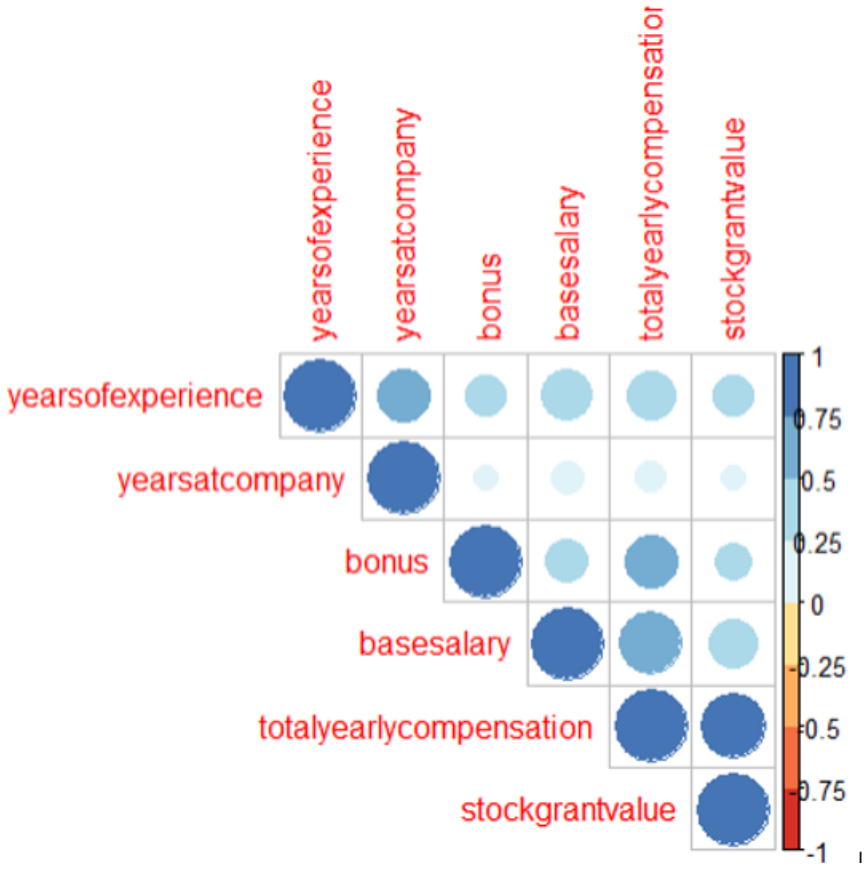


Figure 3-14: Correlation Plots – Salary Data

```
library(Hmisc)
library(xtable)
corstars <-function(numericdata, method=c("pearson", "spearman"), removeTriangle=c("upper", "lower"),
                    result=c("none", "html", "latex")){
    #Compute correlation matrix
    require(Hmisc)
    numericdata <- as.matrix(numericdata)
    correlation_matrix<-rcorr(numericdata, type=method[1])
    R <- correlation_matrix$r # Matrix of correlation coeficients
    p <- correlation_matrix$P # Matrix of p-value
    ## Define notions for significance levels; spacing is important.
    mystars <- ifelse(p < .0001, "****", ifelse(p < .001, "*** ", ifelse(p < .01, "**  ", ifelse(p < .05, "*
", "    "))))
    ## truncatuate the correlation matrix to two decimal
    R <- format(round(cbind(rep(-1.11, ncol(numericdata)), R), 2))[,-1]
    ## build a new matrix that includes the correlations with their apropriate stars
    Rnew <- matrix(paste(R, mystars, sep=""), ncol=ncol(numericdata))
    diag(Rnew) <- paste(diag(R), " ", sep="")
    rownames(Rnew) <- colnames(numericdata)
    colnames(Rnew) <- paste(colnames(numericdata), "", sep="")
    ## remove upper triangle of correlation matrix
    if(removeTriangle[1]=="upper"){
      Rnew <- as.matrix(Rnew)
      Rnew[upper.tri(Rnew, diag = TRUE)] <- ""
      Rnew <- as.data.frame(Rnew)
    }
    ## remove lower triangle of correlation matrix
    else if(removeTriangle[1]=="lower"){
      Rnew <- as.matrix(Rnew)
      Rnew[lower.tri(Rnew, diag = TRUE)] <- ""
      Rnew <- as.data.frame(Rnew)
    }
    ## remove last column and return the correlation matrix
    Rnew <- cbind(Rnew[1:length(Rnew)-1])
    if (result[1]=="none") return(Rnew)
    else{
      if(result[1]=="html") print(xtable(Rnew), type="html")
      else print(xtable(Rnew), type="latex")  }}

corstars(numericdata, result = "none")
```

	totalyearlycompensation <chr>	basesalary <chr>	stockgrantvalue <chr>	bonus <chr>
totalyearlycompensation				
basesalary	0.72****			
stockgrantvalue	0.79****	0.45****		
bonus	0.52****	0.36****	0.26****	
yearsofexperience	0.44****	0.48****	0.33****	0.32****
yearsatcompany	0.18****	0.21****	0.13****	0.12****

Figure 3-15: Correlation Significance Table – Salary Data (P<0.0001 = “****”)

4 MODELING/EVALUATION

Different models were built using different variables and parameters. For this model as per the content learned in the course, we have selected to go with multiple linear regression. As an additional point of the study, we have researched random forest and have executed random forest models as well.

To summarize the following model algorithms were used which are presented in the section below.

1. Multiple Regression Model
2. Random Forrest Regression

4.1 Multiple Regression Model

For multiple regression three models are run with different combinations of dependent variables. The output and the residual plots for all three models are presented in sections below.

4.1.1 Model 1 – Output

The regression model 1 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, tag, and the output is presented in Figure 4-1. From Figure 4-1, the model is not a good fit with the R-Squared value of 0.386, indicating it can only explain 38.6 % of the variation in the data.

```
Call:
lm(formula = totalyearlycompensation ~ title + yearsofexperience +
    yearsatcompany + gender + Race + Education + region + tag,
    data = SalaryData6)

Residuals:
    Min      1Q  Median      3Q     Max
-406135  -50529   -8430   30291 4582875

Residual standard error: 115600 on 11457 degrees of freedom
Multiple R-squared:  0.386,     Adjusted R-squared:  0.3282
F-statistic: 6.686 on 1077 and 11457 DF,  p-value: < 0.00000000000000022
```

Figure 4-1: Multi-Regression Model 1 - Output

4.1.2 Model 1 – Residuals

The regression model 1 model residual plots and the variable residual plots are presented in Figure 4-2 to Figure 4-4. From the residuals, it can be concluded that the model is not a good fit, and we can see that there are many exceptions.

Interpretation of the residuals:

Top left: The residual across the fitted values shows an asymmetrical distribution, with substantial number points on the above y = 0 line, which indicates the prediction is biased and much more salaries were predicted lower than actual.

Top right: the normal Q-Q plot shows is diverging towards the end tail on the higher theoretical quantile side (>2), indicating the deviation from a normal distribution for the errors.

Bottom left: The following scale-location plot indicates the "homoscedasticity" or validity of the 'assumption of equal variance' of the residuals. The red mean line deviates sharply from horizontal toward higher salary side, indicating the equivalence of variances assumption could be violated.

Bottom Right: For model1, a couple of points lie across/ at the Cook's distance reference line.

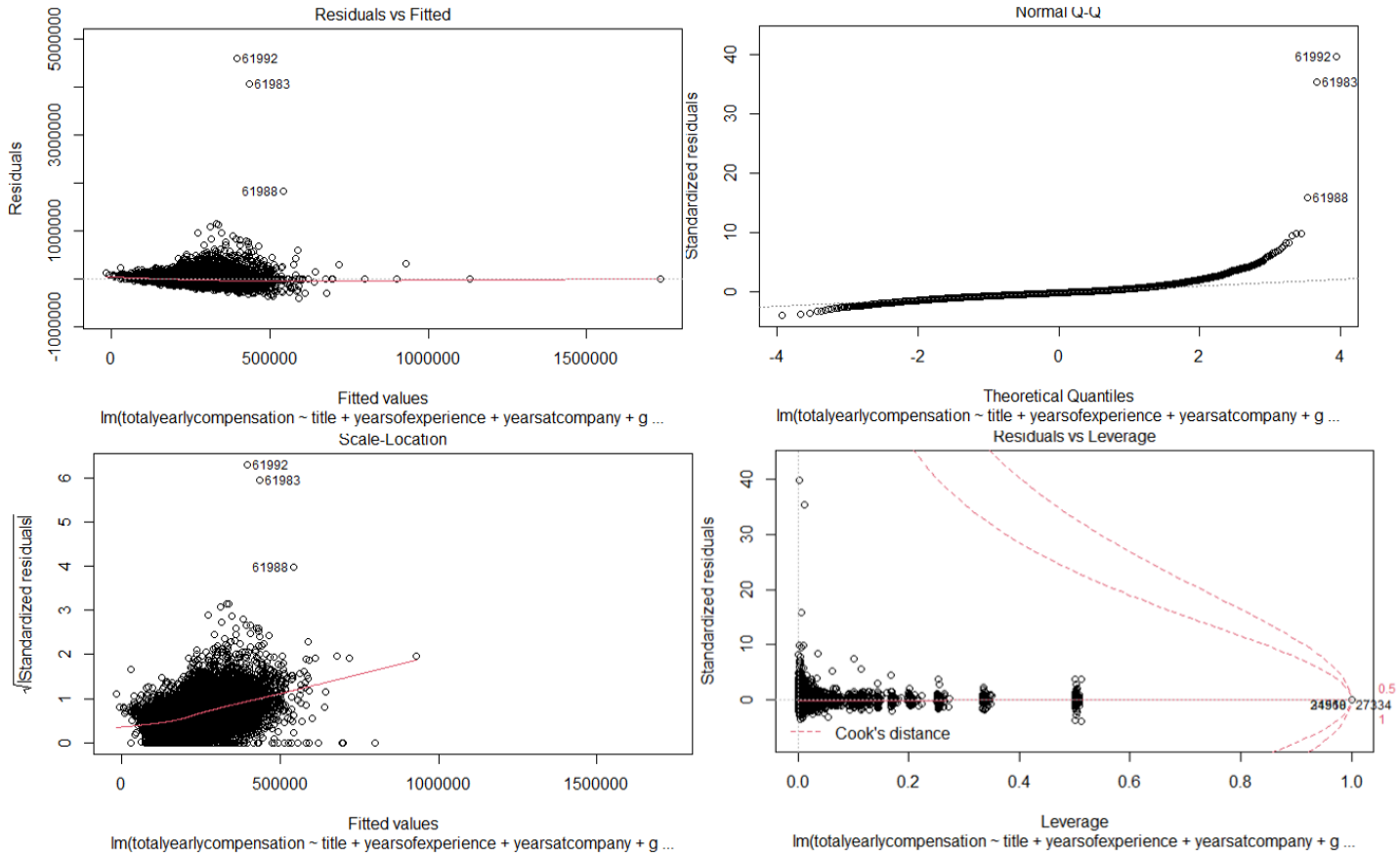


Figure 4-2: Multi-Regression Model 1 – Model Residuals

The individual variable residuals are also densely distributed towards the positive residuals side, indicating a biased prediction (towards lower salary) from the model.

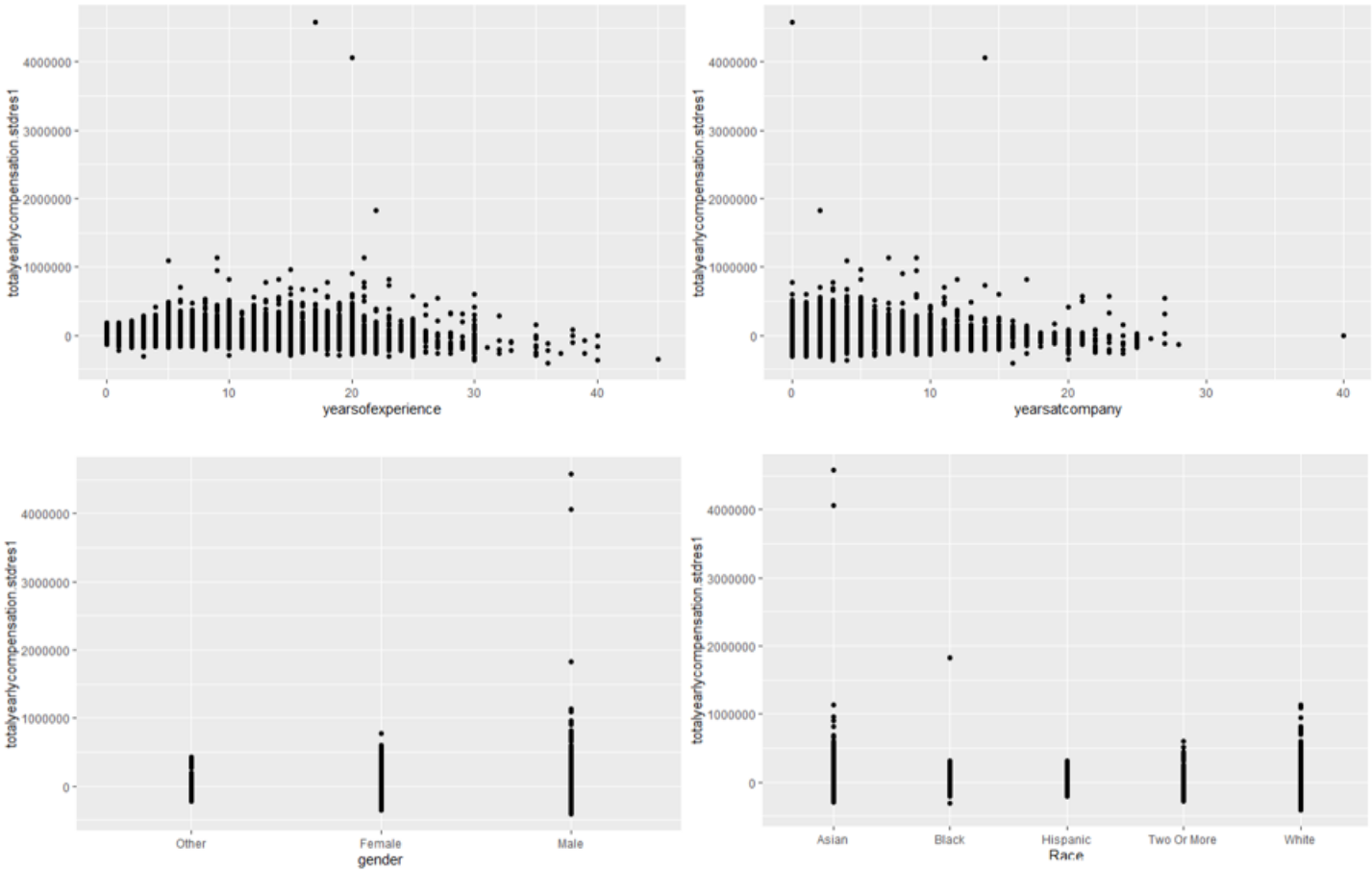


Figure 4-3: Multi-Regression Model 1 – Variable Residuals

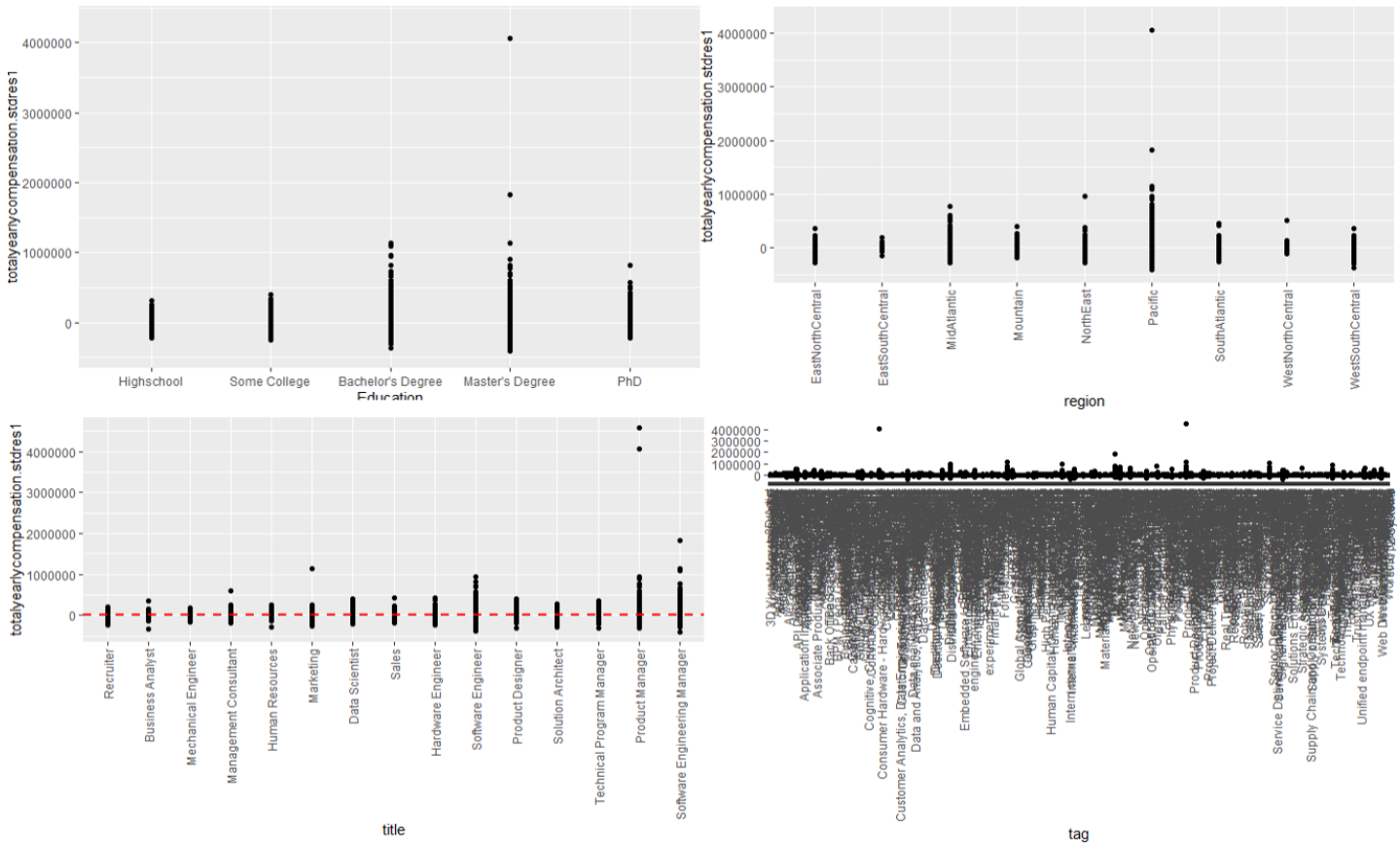


Figure 4-4: Multi-Regression Model 1 – Variable Residuals

4.1.3 Model 2 – Output

The regression model 2 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, and the output is presented in Figure 4-5. From Figure 4-5, the model does not seem to be a very good fit with an R-Squared value of 0.48 (Adj R –Squared 0.475), however, it is better than model 1 in explaining the variation of the data. This model explains 48% of the variation in the data. The F-stat value was 88.99 ($p < 0.00000000000000022$), indicating the model itself is significant.

The numerical variables in model 2, yearsofexperiece and yearsatcompany were highly significant ($p< 0.001$) and multiple subcategories for all the categorical variables in the model were found to be statistically significant ($p< 0.05$). For example, in the 'region' category, the pacific and Mid-Atlantic were Statistically highly significant ($p< 0.001$).
Since all the variables were found statistically significant and the R-squared value was not very high, we did not run the model with the reduced number of variables. However, as seen from section 3.6 (univariate analysis histograms), the histogram of log(totalcompensation) is much more normally distributed, so next (section 4.1.5), we ran the linear regression using the log of totalcompensation variable.

```
Call:
lm(formula = totalyearlycompensation ~ company + title + yearsofexperience +
    yearsatcompany + gender + Race + Education + region, data = SalaryData6)

Residuals:
    Min      1Q  Median      3Q     Max
-347745  -43501   -5936   29835 4497052

Residual standard error: 102200 on 12405 degrees of freedom
Multiple R-squared:  0.4806,    Adjusted R-squared:  0.4752
F-statistic: 88.99 on 129 and 12405 DF,  p-value: < 0.00000000000000022
```

Figure 4-5: Multi-Regression Model 2 - Output

4.1.4 Model 2 – Residuals

The regression model 2 model residual plots and the variable residual plots are presented in Figure 4-6 to Figure 4-8. From the residuals, it can be concluded that the model is a fair fit.
Interpretation of the residuals:
Top left: The residual across the fitted values shows a slightly asymmetrical distribution, with more points on above y = 0, which indicates the prediction is biased and more salaries were predicted lower than actual.
Top right: the normal Q-Q plot shows is diverging towards the end tail on the higher theoretical quantile side (>2), indicating the deviation from the normal distribution.
Bottom left: The following scale-location plot indicates the "homoscedasticity" or validity of the 'assumption of equal variance' of the residuals. The red mean line deviates from horizontal toward higher salary side. No drastic deviation was observed, it could be inferred that the assumption of equal variance is not violated.
Bottom Right: For model2, no influential points lie across the Cook's distance. There are a couple of some observations (61992, 61983) that are away from the residual cluster; however, they are still well below Cook's distance limit (dashed red line).

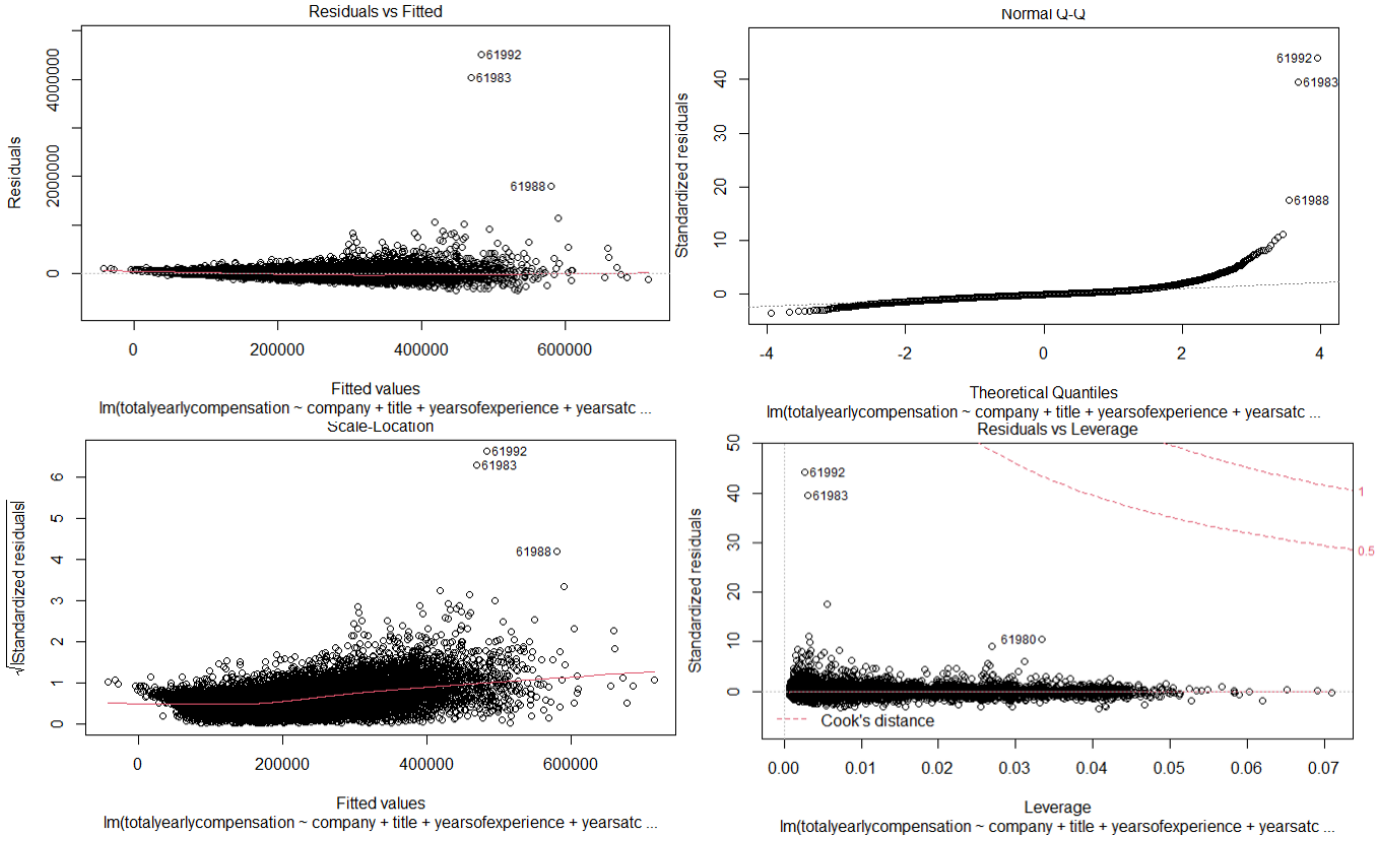


Figure 4-6: Multi-Regression Model 2 – Model Residuals

Residuals of individual variables
The residuals for the variables as shown below are denser and more spread (distant) on the positive side from the y = 0 axis (shown by the dashed lines), which indicates that the model predicts salaries lower than the actual salaries.

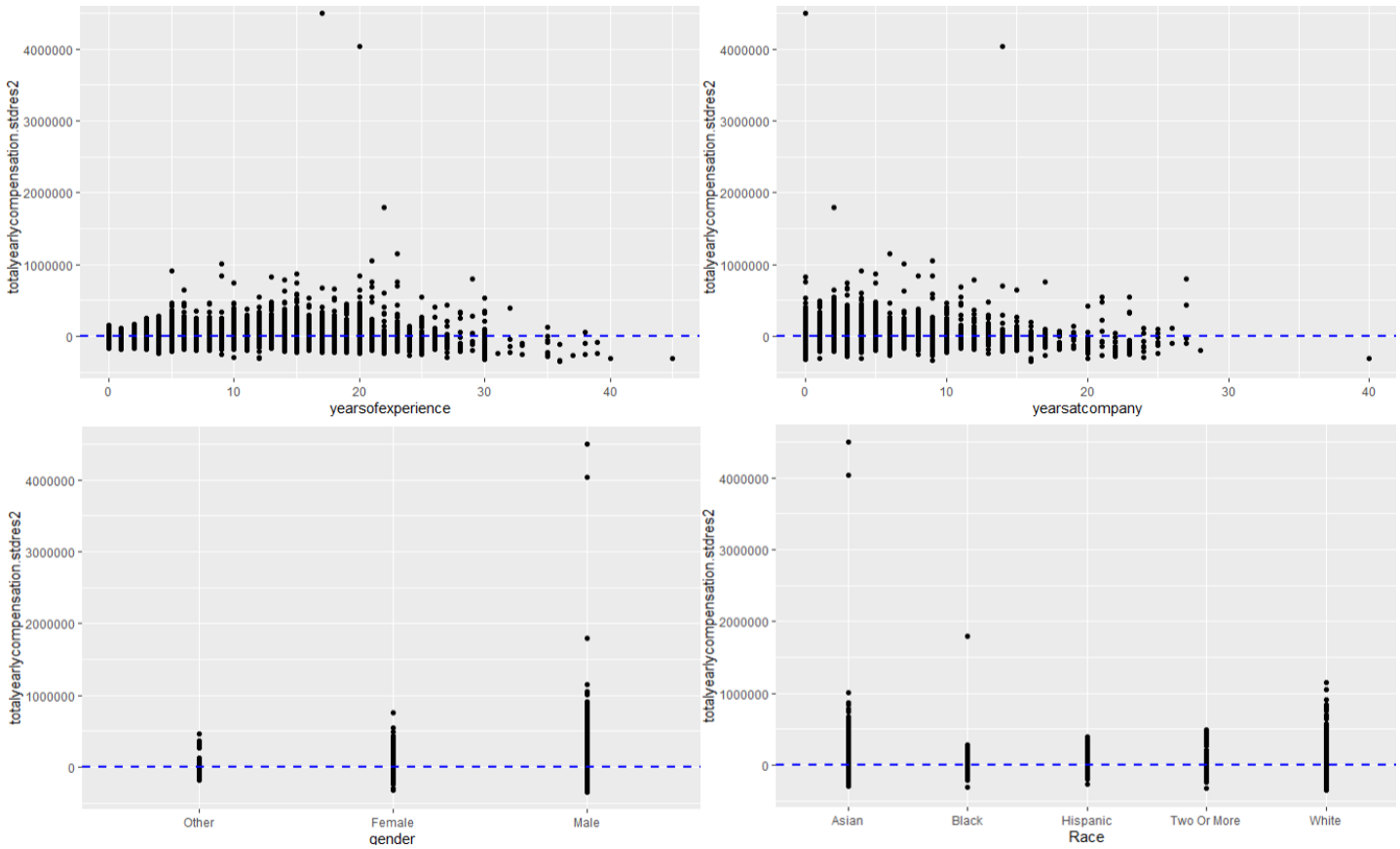


Figure 4-7: Multi-Regression Model 2 – Variable Residuals

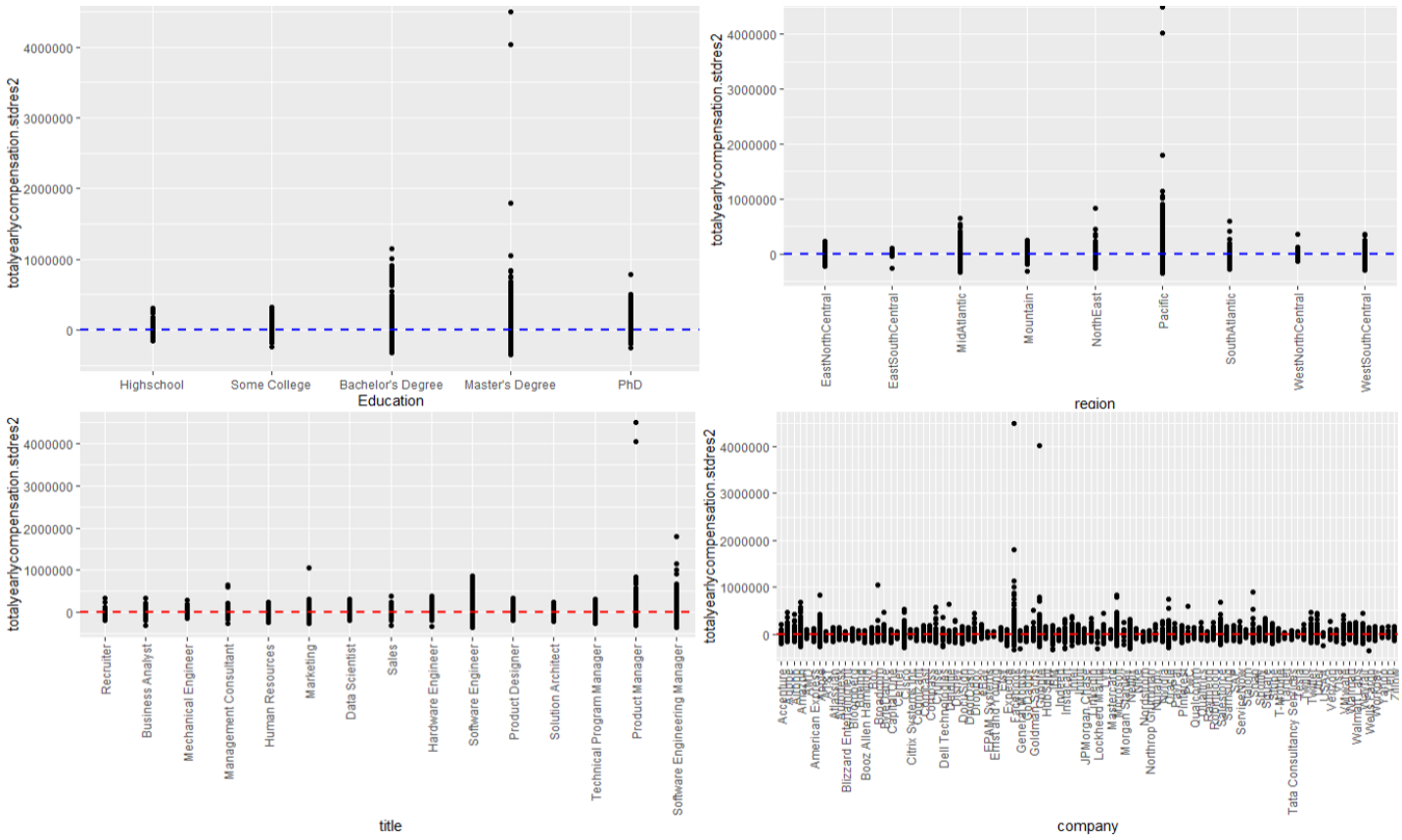


Figure 4-8: Multi-Regression Model 2 – Variable Residuals

4.1.5 Model 3 (Log Model) – Output

The regression model 3 was executed with log value of y variable and the following x variables in consideration; title, yearsofexperience, yearsatcompany, company, gender, Race, Education, region, and the output is presented in Figure 4-9. The model is a better fit with an R-Squared value of 0.681 (adj R Squared = 0.678). This means it explains 68% of the variance in the data. The F-stat value was 205.6 ($p < 0.00000000000000022$). The higher R squared and F-statistics value of model 3 as compared to model 1 and model 2 indicates, it is better in explaining the variation of the dataset with the model variables than models 1 and 2.

For model3 as well, the numerical variables yearsofexperiece and yearsatcompany were highly statistically significant ($p< 0.001$), and multiple subcategories for all the categorical variables in the model were found to be statistically significant ($p< 0.05$). For example, all the categories in the job 'title' variable were statistically highly significant ($p< 0.001$).

```
Call:
lm(formula = logTotalSalary ~ company + title + yearsofexperience +
    yearsatcompany + gender + Race + Education + region, data = SalaryData6)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0685 -0.1541 -0.0097  0.1502  2.3115

Residual standard error: 0.267 on 12405 degrees of freedom
Multiple R-squared:  0.6814,    Adjusted R-squared:  0.678
F-statistic: 205.6 on 129 and 12405 DF,  p-value: < 0.00000000000000022
```

Figure 4-9: Multi-Regression Model 3 - Output

4.1.6 Model 3 – Residuals

The regression model 3 model residual plots and the variable residual plots are presented in Figure 4-10 to Figure 4-12. From the residuals, it can be concluded that the model is a better fit. The model is better than models 1 & 2 and is the most parsimonious of all.

Interpretation of the residuals:

Top left: The residual across the fitted values shows a symmetrical distribution. The distribution is not curved, it has equally spread residuals around the horizontal line without any noticeable distinct pattern, indicating that it is not a non-linear relationship.

Top right: the normal Q-Q plot the normal Q-Q plot shows a fairly good alignment to the dotted straight line in the quantile region (-2,2) showing closeness to a normal distribution of the errors in that region, the plot diverges towards the end tail on both sides.

Bottom left: The following scale-location plot indicates the "homoscedasticity" or validity of the 'assumption of equal variance' of the residuals. The red mean line deviates from horizontal toward higher salary side. No drastic deviation was observed, it could be inferred that the assumption of equal variance is not violated.

Bottom Right: For model3, no influential points lie across the Cook's distance.

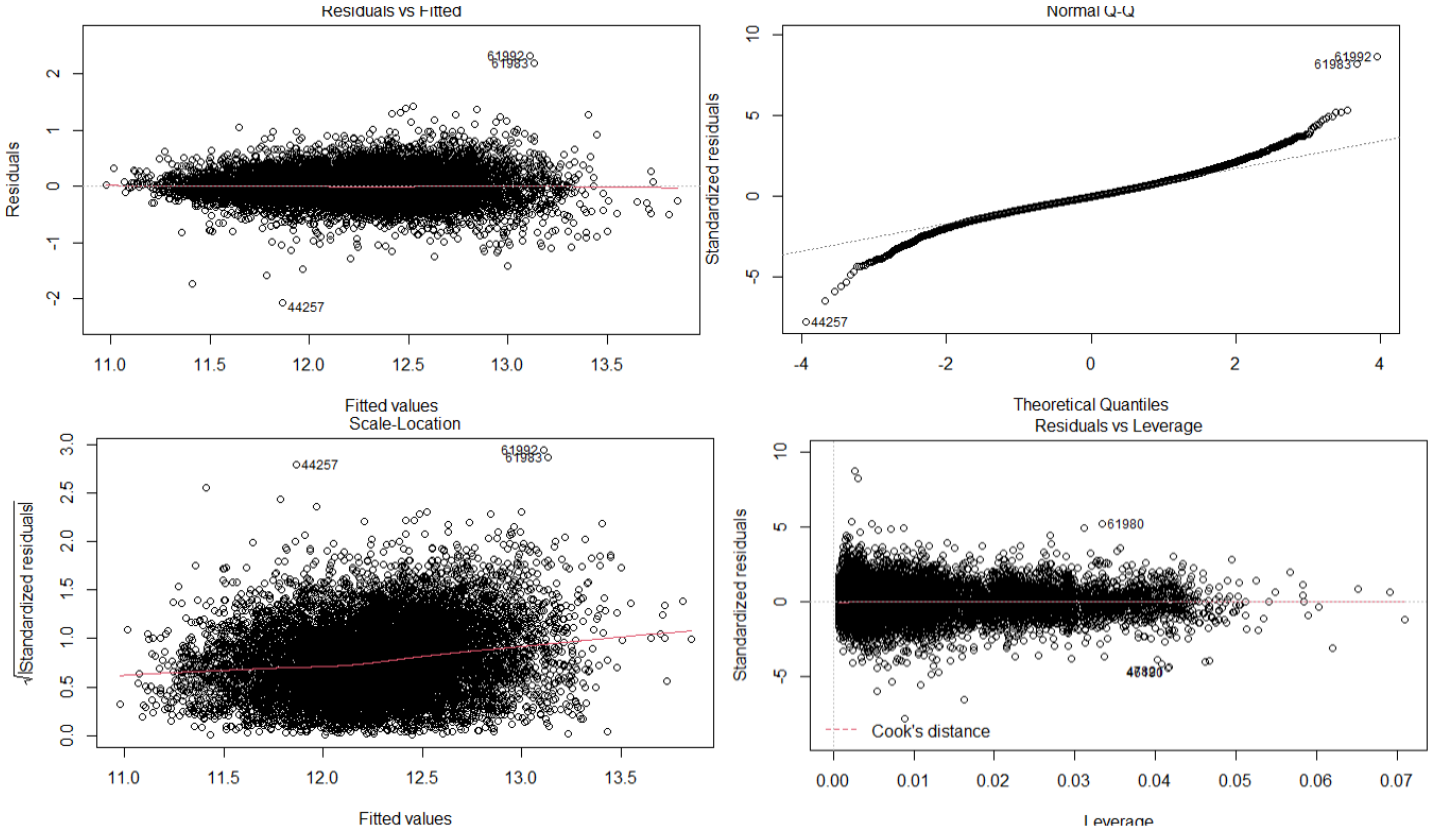


Figure 4-10: Multi-Regression Model 3 – Model Residuals

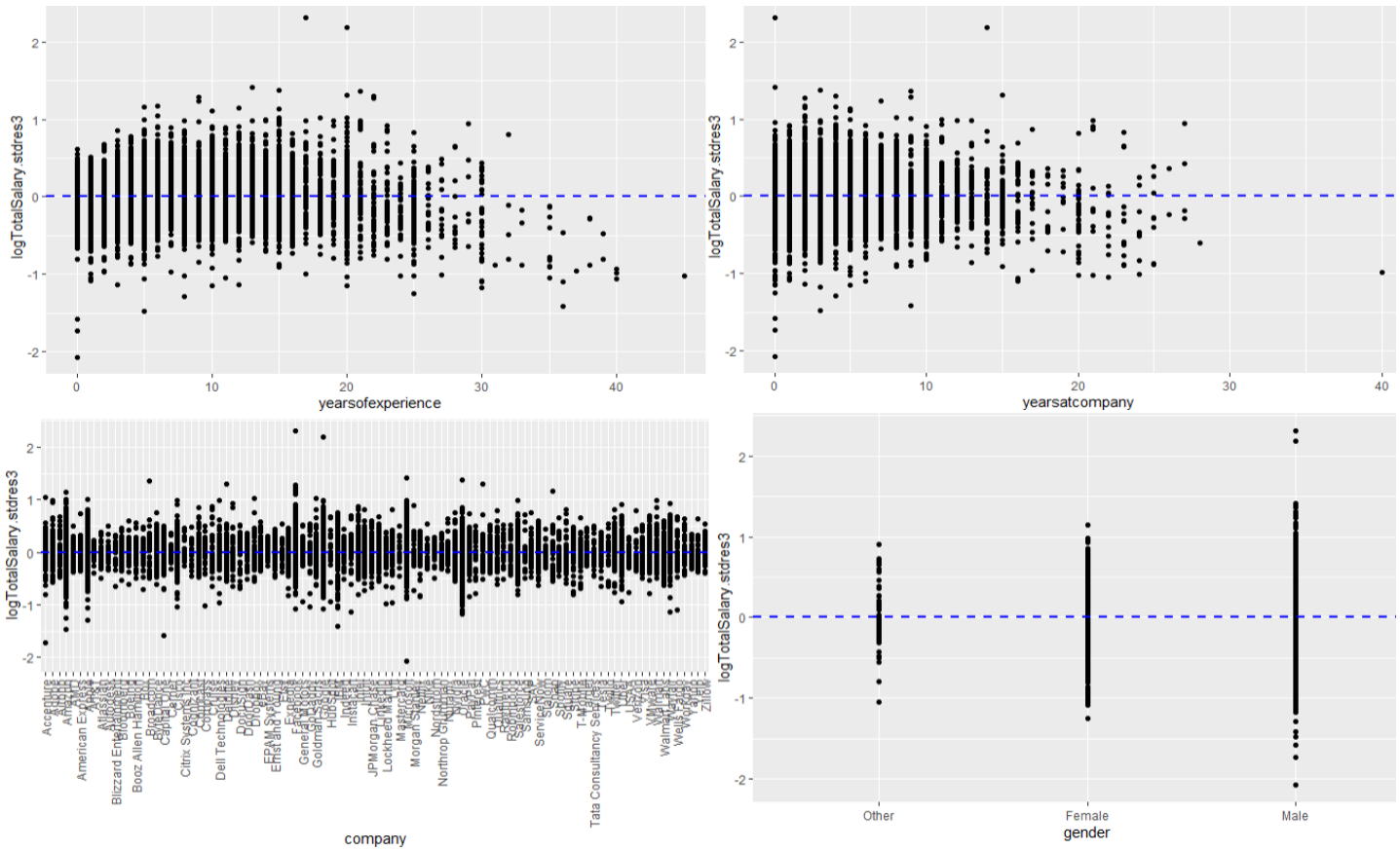


Figure 4-11: Multi-Regression Model 3 – Variable Residuals

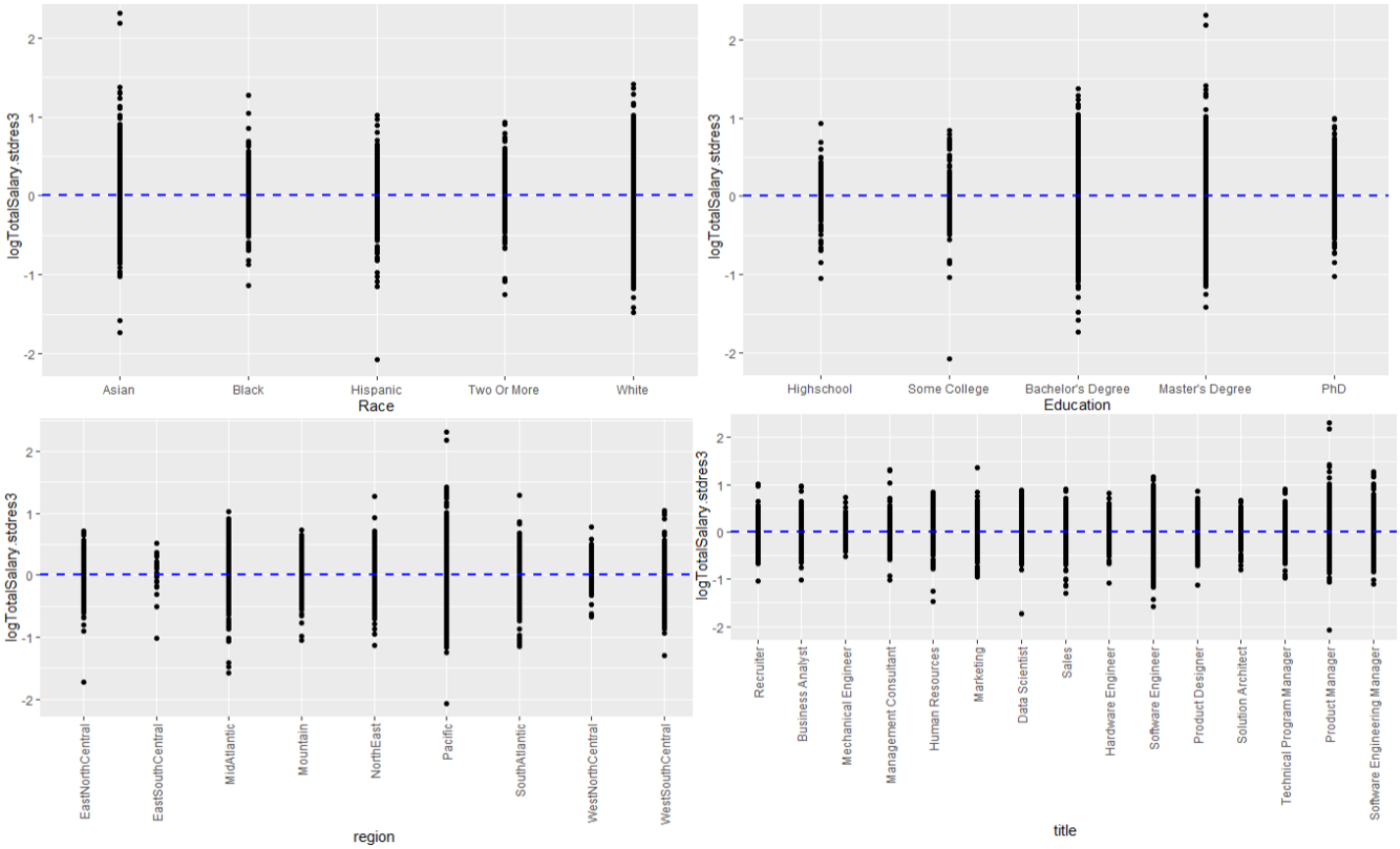


Figure 4-12: Multi-Regression Model 3 – Variable Residuals

4.1.7 Test for collinearity (test with Variance Inflation Factor, VIF)

Collinear variables inflate the model, because they are likely are explaining the same variance in the model and can result in an apparent better fit, rather than the real fit. Also, since collinear variables compete with each other sometimes, neither of them shows statistical significance in the model, even though they might be. We used the 'VIF' function in the 'car' package for this test.

Models 2 and 3 were checked for collinearity using VIF regression (models 2 and 3 have the same independent variables). A GVIF value (proportional change of the standard error and confidence interval of their coefficients due to the level of collinearity) of greater than 5, would indicate collinearity between the variables. The variables 'company 'and 'region' had a very high GVIF value, much greater than 5, therefore we can interpret that there is collinearity of these variables with other variables. An accurate contribution of these variables in the model is hence difficult to assess. It is also possible that the two variables 'company' and 'region' are highly correlated with each other. One possibility could be that a large number of software/ IT companies are located in the pacific region and also has higher salaries than several other regions.[6]. The results of the collinearity test are presented in Table 4-1.

##	GVIF	Df	GVIF^(1/(2*Df))
## company	44.449976	95	1.020171
## title	4.724101	14	1.057019
## yearsofexperience	1.765986	1	1.328904
## yearsatcompany	1.629641	1	1.276574
## gender	1.154667	2	1.036607
## Race	1.375968	4	1.040701
## Education	1.412016	4	1.044071
## region	11.210974	8	1.163062

Table 4-1: Collinearity - VIF

4.2 Random Forest Model

For random forest three models are run with different combinations of dependent variables. The output and the prediction/variable significance plots for all three models are presented in the sections below.

4.2.1 Model 1 – Output

The random forest model 1 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, tag, and the output is presented in Figure 4-13. From Figure 4-13 the model is not a good fit with a % variance explained value of 33.39.

```
Call:
 randomForest(formula = totalyearlycompensation ~ title + yearsofexperience +
yearsatcompany + gender + Race + Education + region + tag,      data = SalaryData6,
mtry = 3, importance = TRUE, proximity = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 13254805243
                    % Var explained: 33.39
```

Figure 4-13: Random Forest Model 1 - Output

4.2.2 Model 1 – Variable Significance / Predicted Plots

The random forest model 1 variable significance and the predicted plot are presented in Figure 4-2 to Figure 4-4. From the predicted vs actual plot, the model is not a good fit.

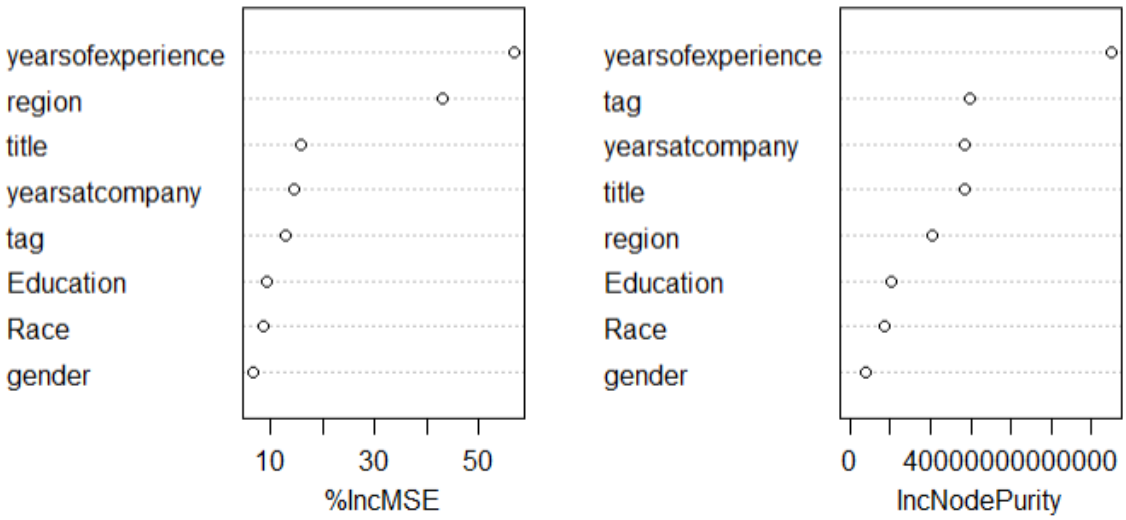


Figure 4-14: Random Forest Model 1 – VariableSignificance

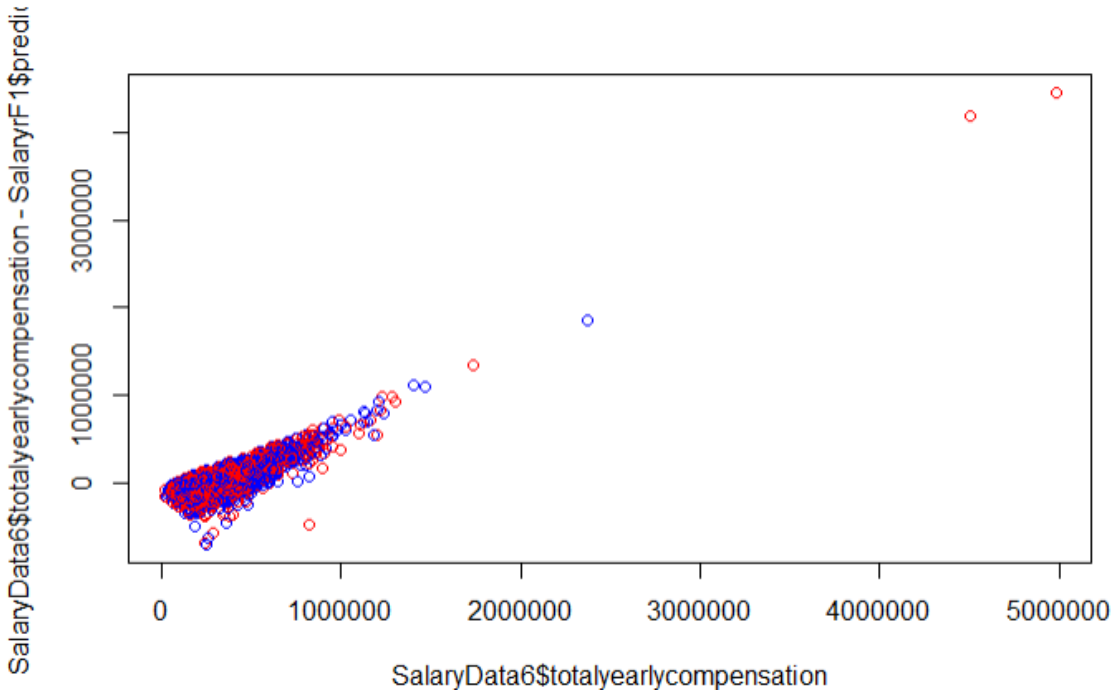


Figure 4-15: Random Forest Model 1 – Predicted vs Actual Plot

4.2.3 Model 2 – Output

The random forest model 2 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, and the output is presented in Figure 4-16. From Figure 4-16, the model is a better fit compared to model 1 with a % variance explained value of 46.09.

```
Call:
 randomForest(formula = totalyearlycompensation ~ company + title +
yearsofexperience + yearsatcompany + gender + Race + Education +      region, data =
SalaryData6, mtry = 3, importance = TRUE,      na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 10727993619
                    % Var explained: 46.09
```

Figure 4-16: Random Forest Model 2 - Output

4.2.4 Model 2 – Variable Significance / Predicted Plots

The random forest model 2 variable significance and the predicted plot are presented in Figure 4-17 to Figure 4-18. From the predicted vs actual plot, the model is ok to fit, and we can see that the predictions are all over the place.

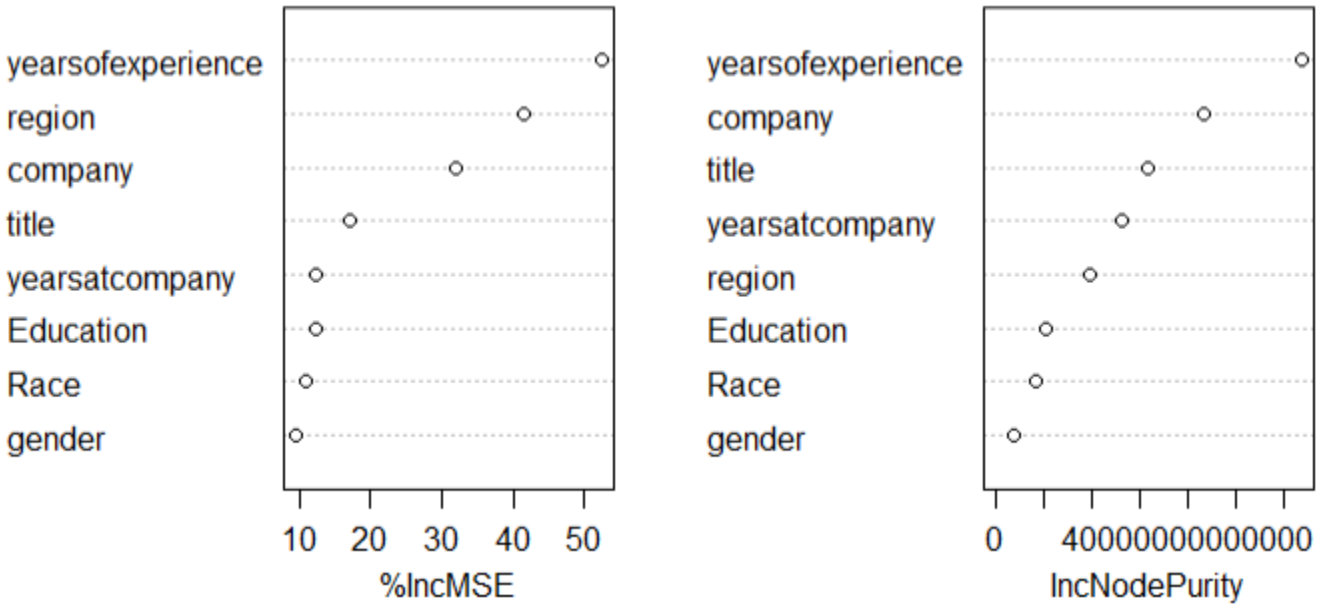


Figure 4-17: Random Forest Model 2 – Variable Significance

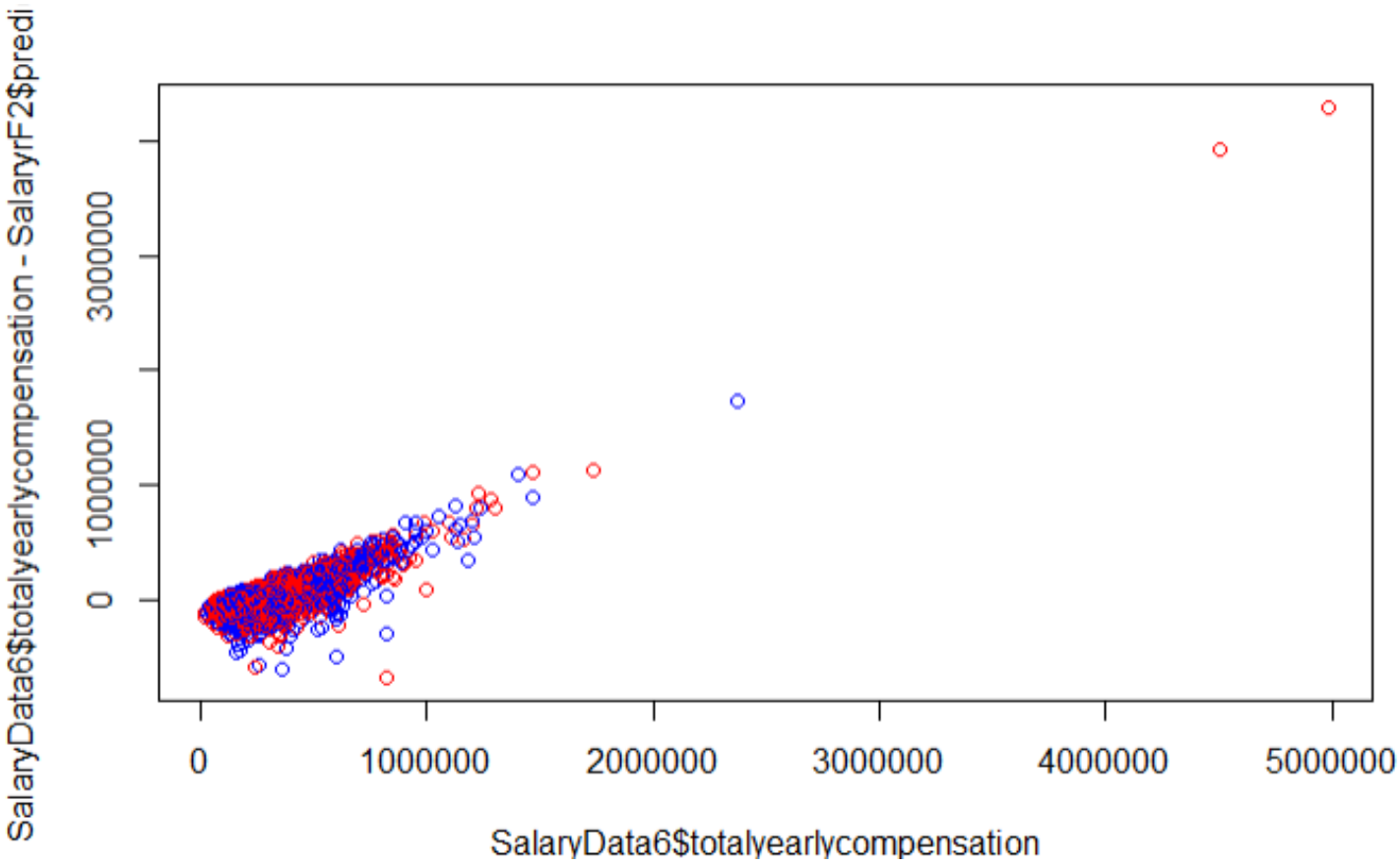


Figure 4-18: Random Forest Model 2 – Predicted vs Actual Plot

4.2.5 Model 3 (Log) – Output

The random forest model 3 was executed by taking the log of y variable and the following x variables into consideration; title, yearsofexperience, yearsatcompany, company, gender, Race, Education, region, and the output is presented in Figure 4-19. From Figure 4-19, the model is a good fit with a % variance explained value of 62.99.

```
Call:
 randomForest(formula = logTotalSalary ~ company + title + yearsofexperience +
yearsatcompany + gender + Race + Education + region, data = SalaryData6,      mtry = 3,
importance = TRUE, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.08279153
                    % var explained: 62.6
```

Figure 4-19: Random Forest Model 3 - Output

4.2.6 Model 3 – Variable Significance / Predicted Plots

The random forest model 3 variable significance and the predicted plot are presented in Figure 4-20 to Figure 4-21. From the predicted vs actual plot, the model is an ok fit compared to model 2 and we can see that the predictions are all over the place.

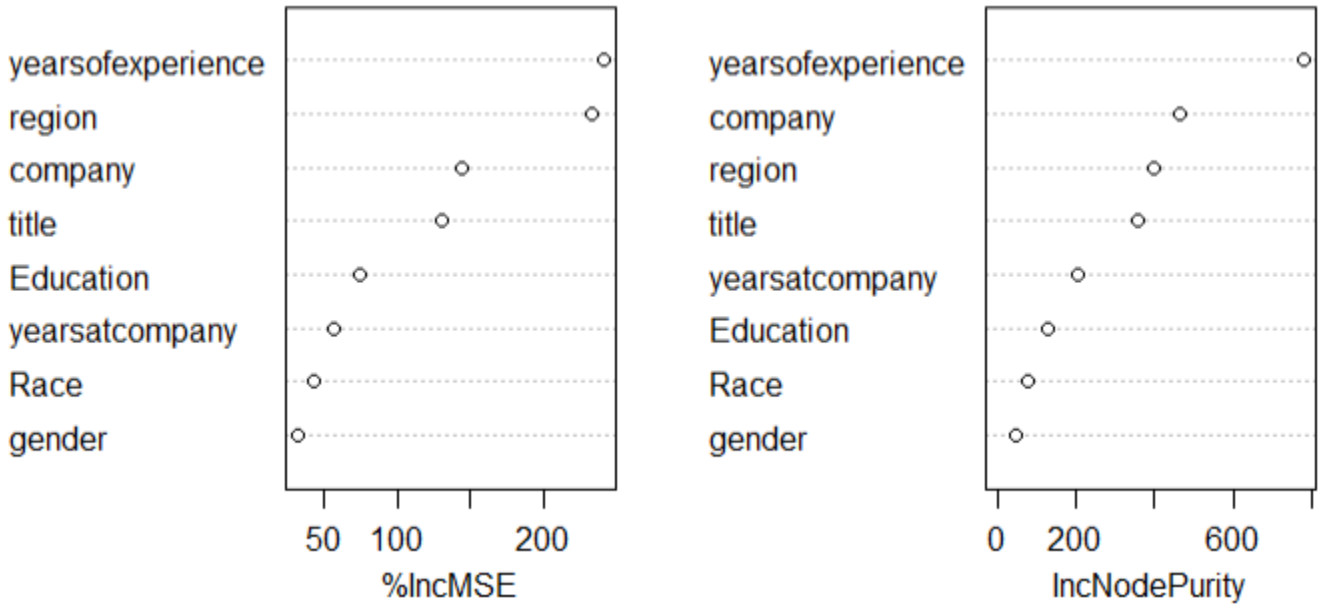


Figure 4-20: Random Forest Model 3 – Variable Significance

Figure 4-21: Random Forest Model 3 – Predicted vs Actual Plot

5 RESULTS

The results from the three models and the two different regression algorithms are summarized and presented in Table 5-1 below. From the result, the most parsimonious model is model 3, a linear regression model with an R squared value of 0.68. (Explains 68% of the variance in the data)

Model	# Features	List of Actual n Features Selected	R-Squared	Adj R-Squared	P	% Variance Explained
Linear Regression Model1	8	title , yearsofexperience , yearsatcompany , gender, Race , Education , region, tag	0.386	0.328	2.20E-16	
Linear Regression Model2	8	title , yearsofexperience , yearsatcompany , gender , Race , Education , region, company	0.481	0.475	2.20E-16	
Linear Regression Model3 (log)	8	title , yearsofexperience , yearsatcompany , gender , Race , Education , region, company	0.681	0.678	2.20E-16	
Random Forest Model1	8	title , yearsofexperience , yearsatcompany , Race , Education , region, tag				33.39
Random Forest Model2	8	title , yearsofexperience , yearsatcompany , gender , Race , Education , region, company				46.09
Random Forest Model3 (log)	8	title , yearsofexperience , yearsatcompany , gender , Race , Education , region, company				62.60

Table 5-1: Model Results

6 LIMITATIONS

One of the key limitations we had was that we were predicting numerical value dependent on many categorical variables and as such running the code took a lot of time. There was also a strong correlation between bonus, base salary, stockgrants, and the total yearly compensation which has introduced bias into our model.

Another limitation was quite a bit of null values and though the data was acquired from levels. fyi which is mostly accurate, we had a few entries where people reported quite conflicting salaries due to which we had few outliers and exceptions in the model.

7 CONCLUSION/RECOMMENDATIONS

To conclude we say that the top FAANG / MAMAA companies tend to be on the higher spectrum salaries, and we identified that the top 20 companies tend to pay more salaries. From Figure 7-1 & Figure 7-2, we recommend applying to the top 25 companies which pay higher salaries than others. We also recommend checking the years of experience to understand what salary they can negotiate for.

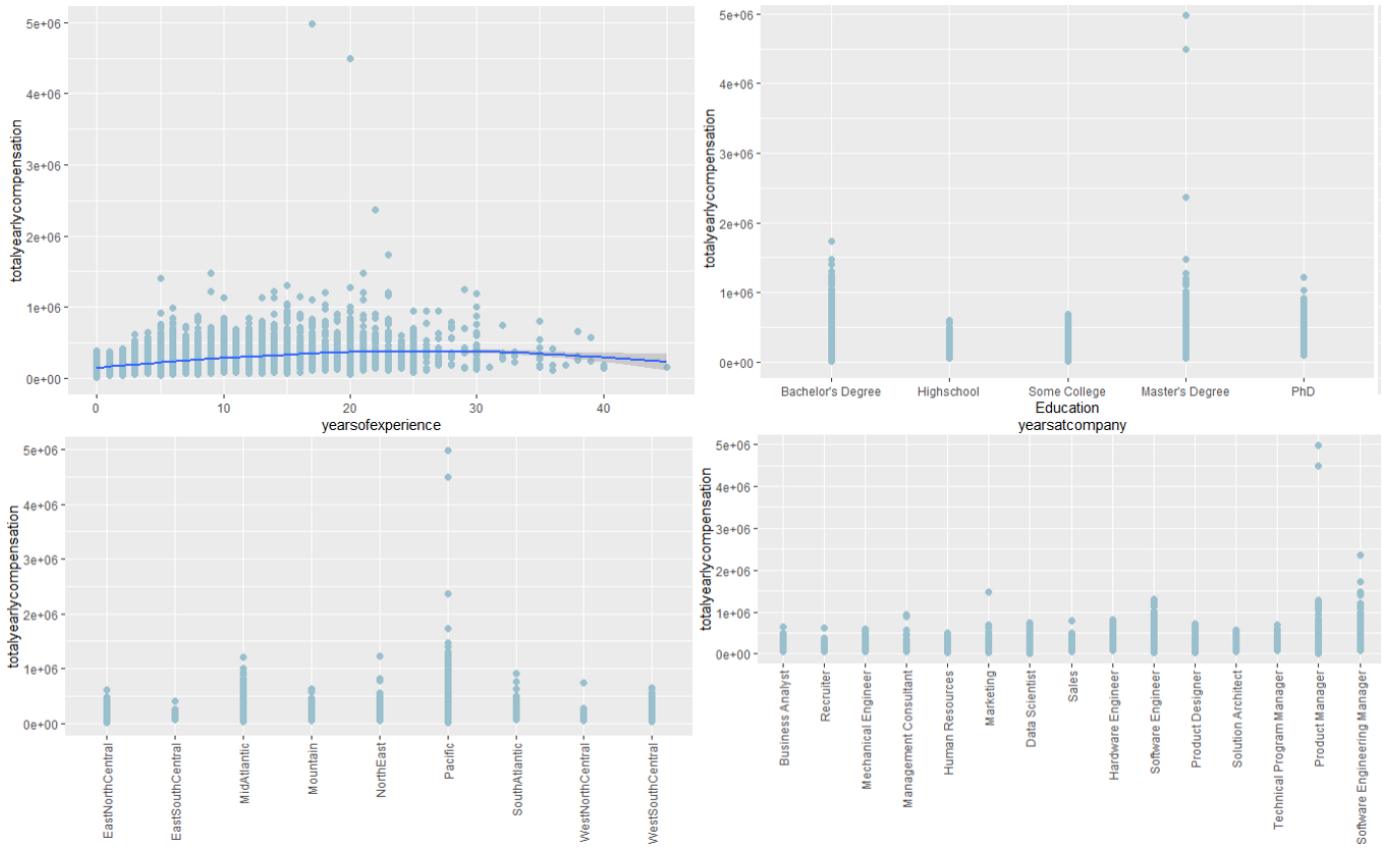


Figure 7-1: Yearlcompensations Vs Deciding Factors

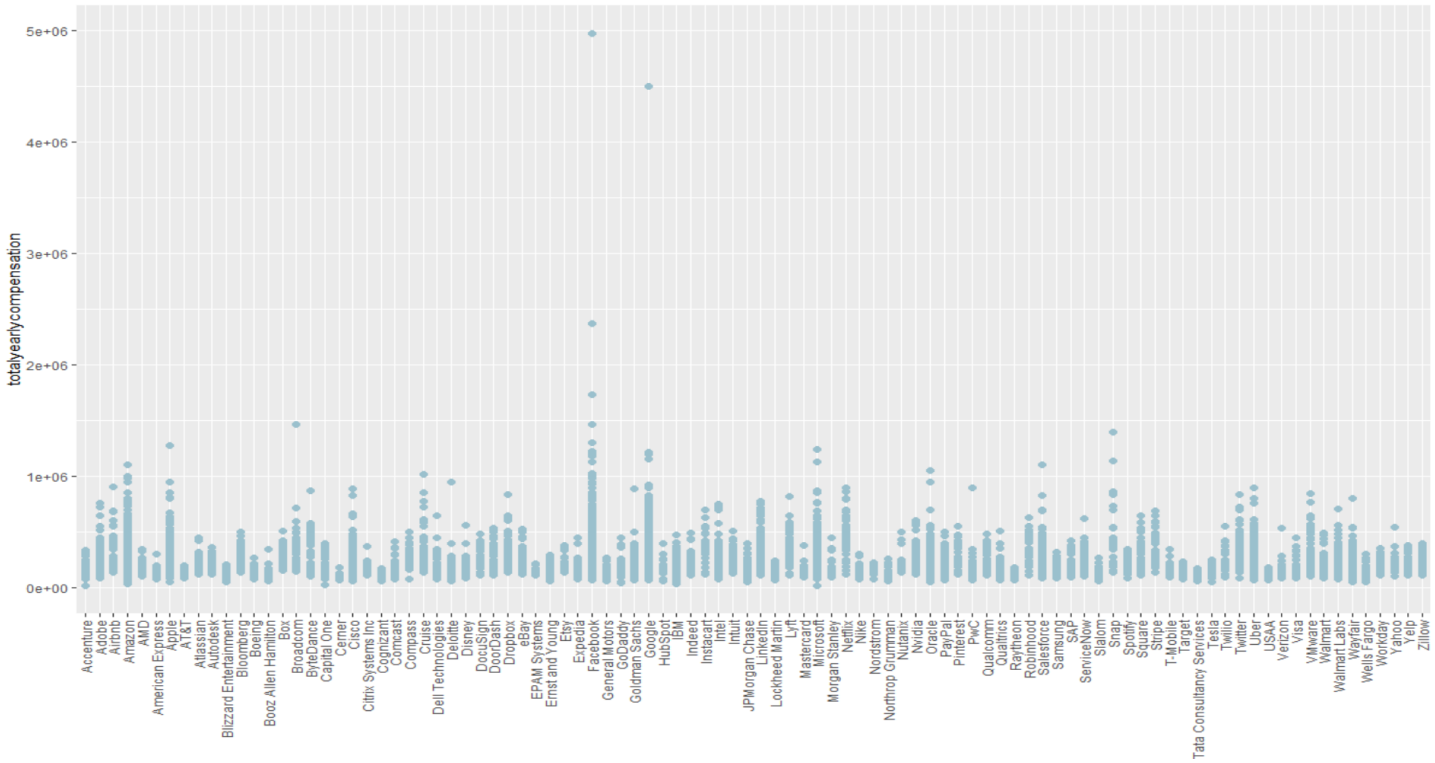
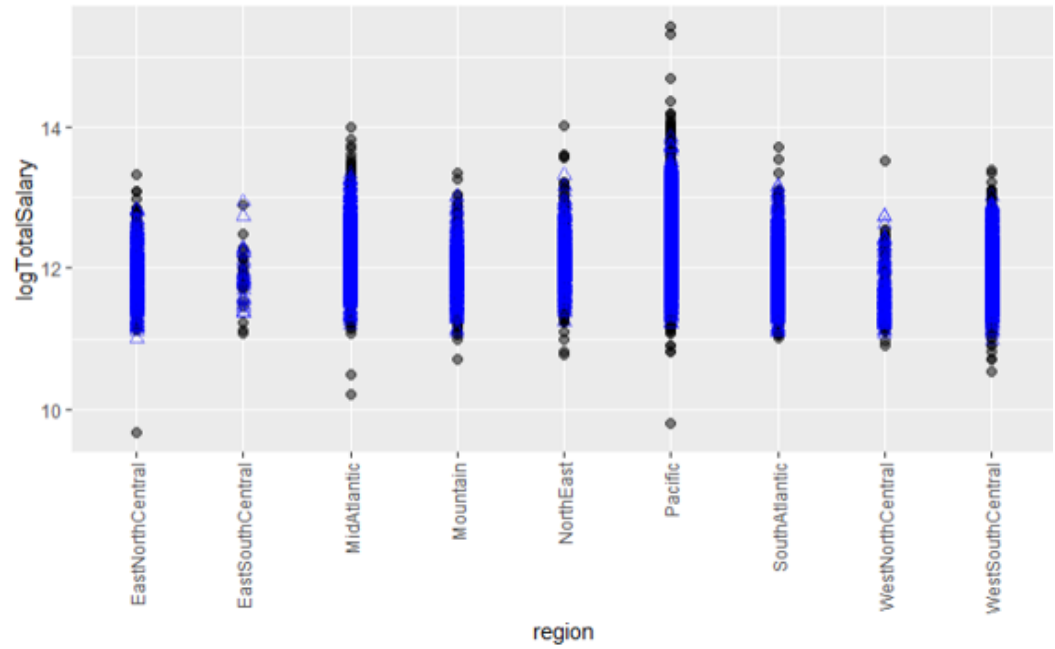


Figure 7-2: Yearlcompensations Vs Top Companies

8 PREDICTIONS – MOST PARSIMONIOUS MODEL

From Table 5-1 the most parsimonious model is established as model 3 with linear regression. The prediction algorithm is run for this model and the predicted and actual values are plotted for each variable. The plots are presented in Figure 8-1 to Figure 8-4.

It can be seen from the following plots that the model is a good fit for the average salaries and does not predict the higher or lower ranges of the salary well. One reason could be that the number of observations (ref sec 3.6, histogram) in the lower/ higher end of the range is lower in number.



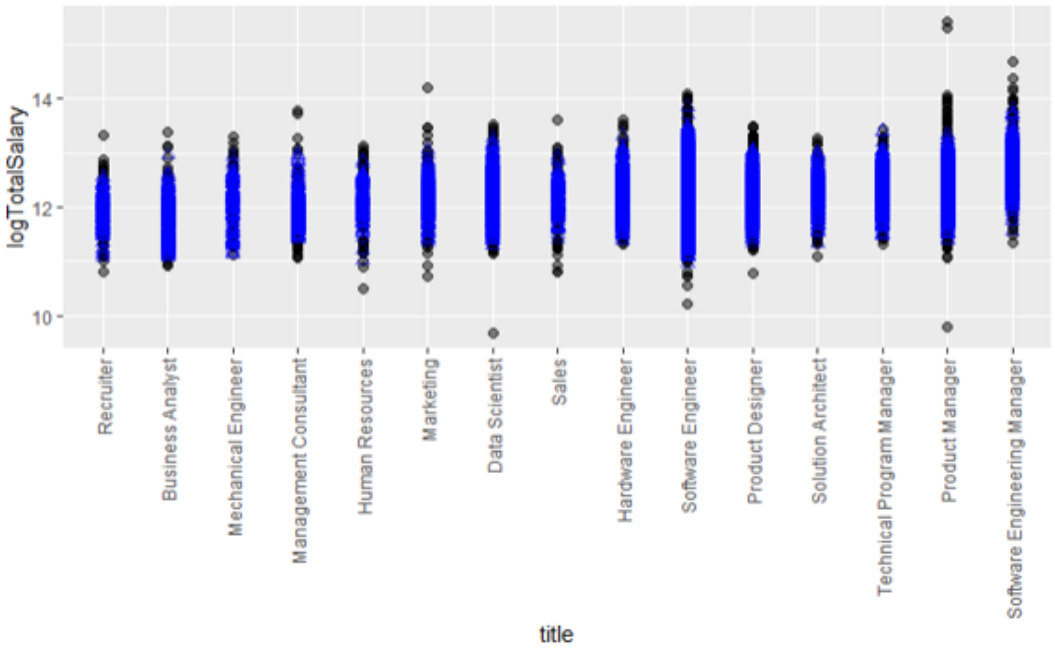
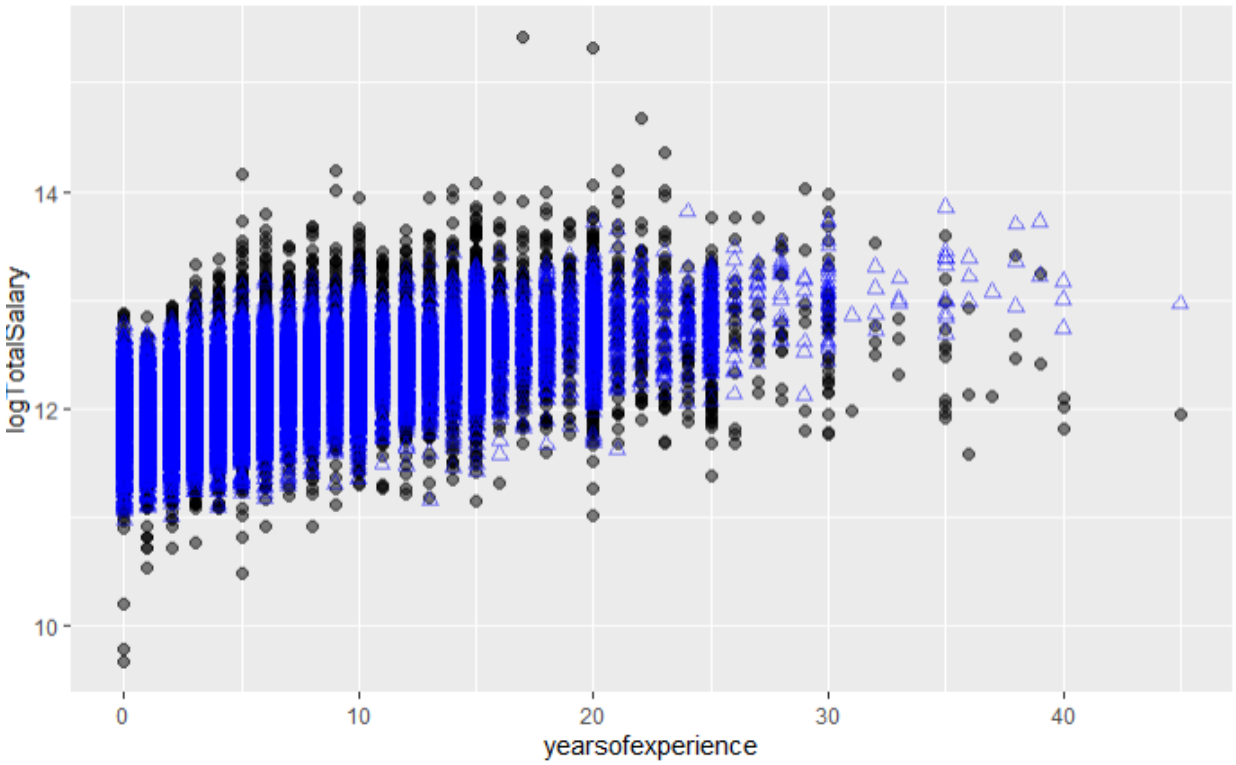


Figure 8-1: Predicted Vs Actual – Parsimonious Model



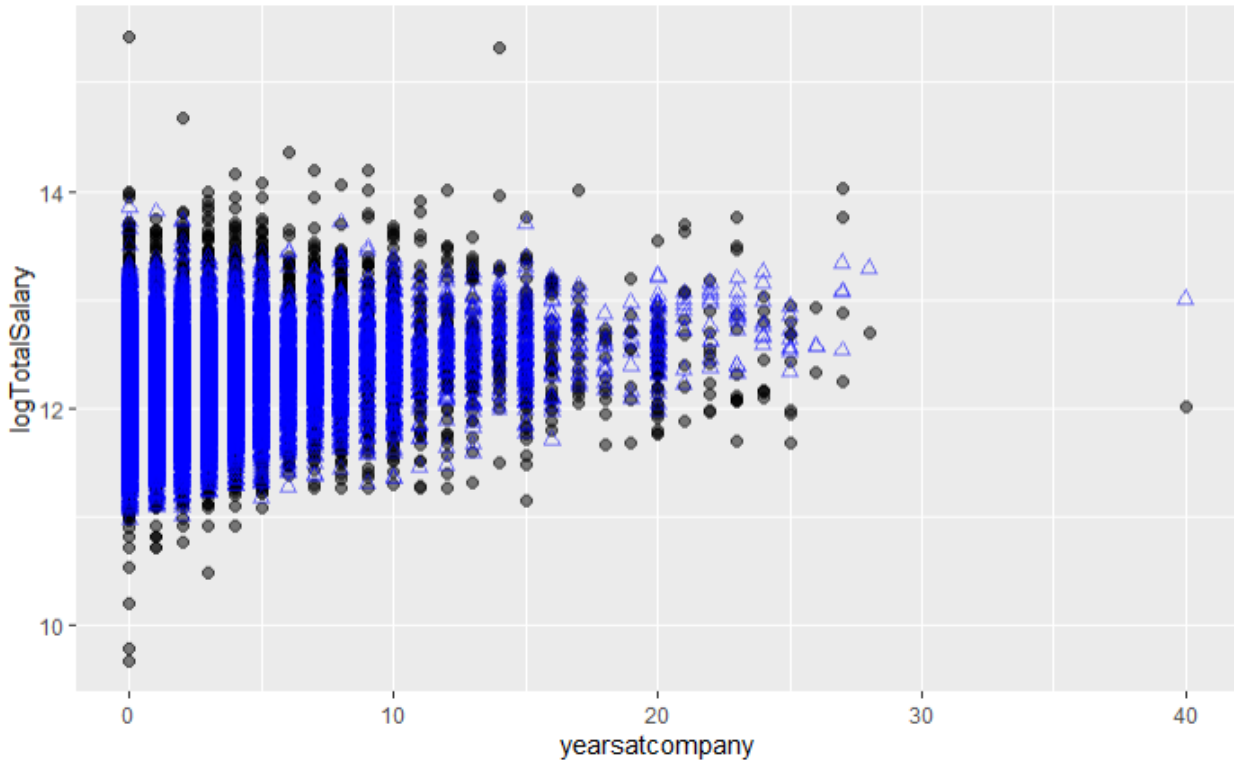


Figure 8-2: Predicted Vs Actual – Parsimonious Model – Cont…

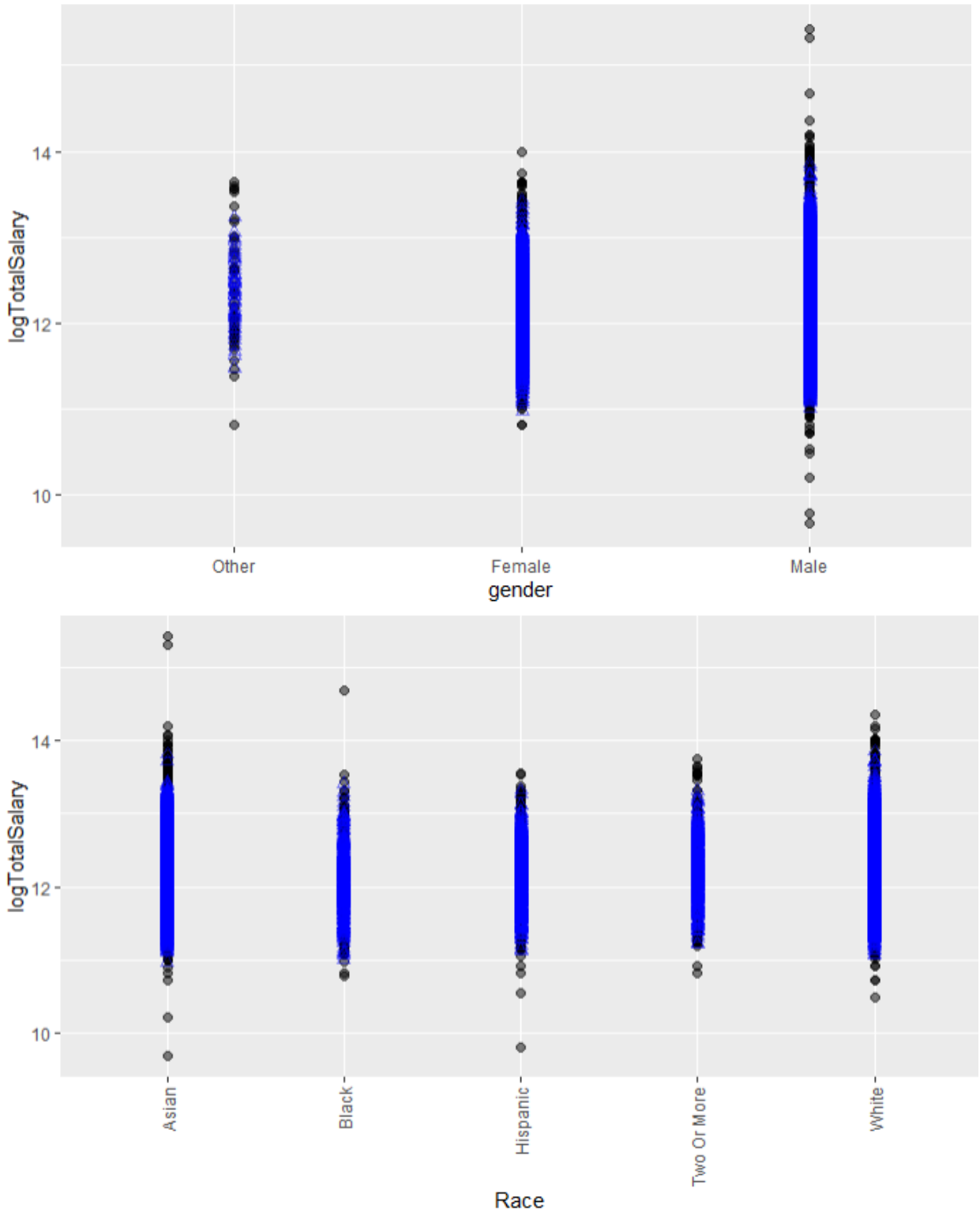


Figure 8-3: Predicted Vs Actual – Parsimonious Model – Cont…

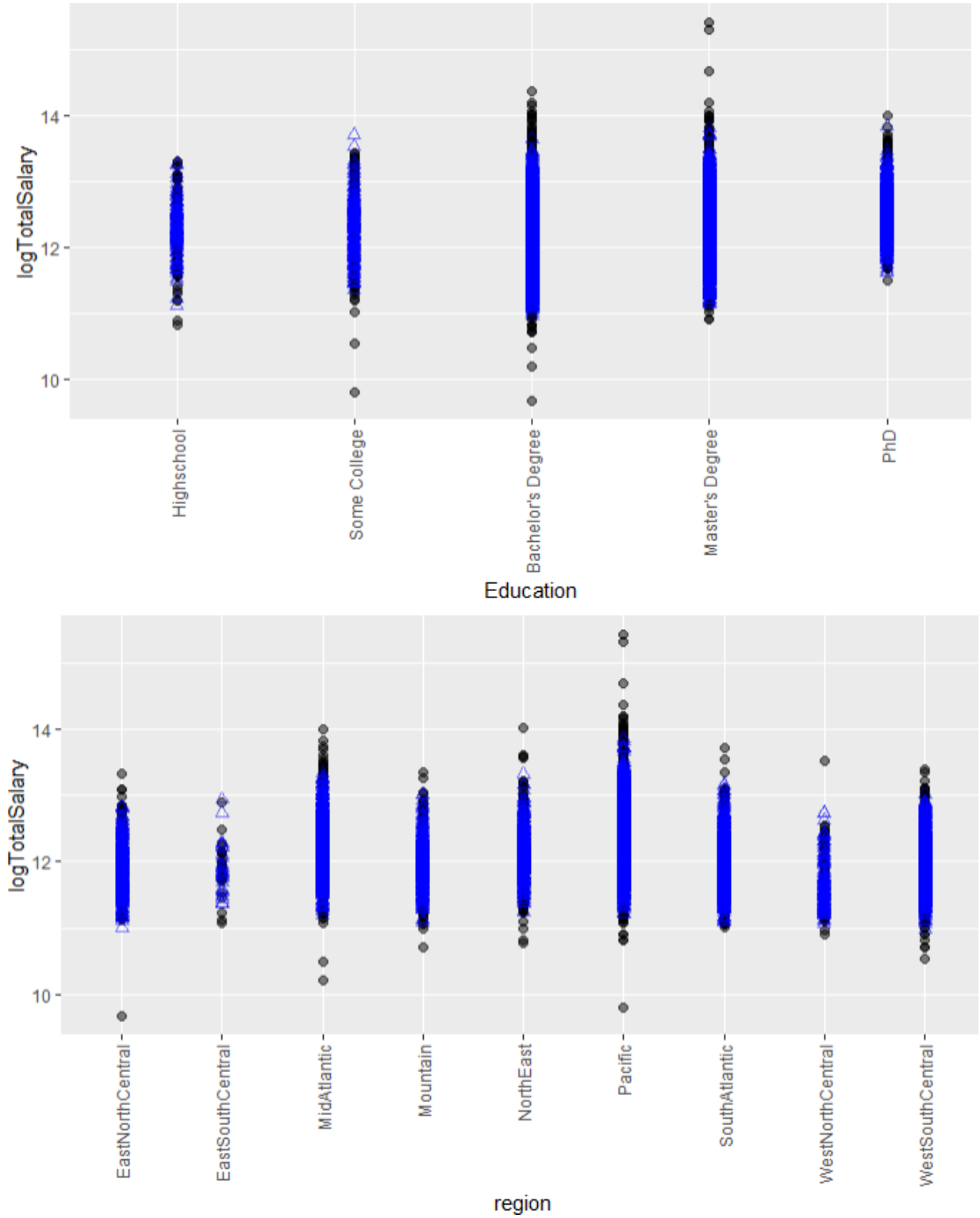


Figure 8-4: Predicted Vs Actual – Parsimonious Model – Cont…

9 REFERENCES

[1] https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries

[2] https://www.levels.fyi/comp.html?track=Software%20Engineer&region=819

[3] https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c

[4] https://view.genial.ly/608626ff91d129103155b087/interactive-content-311-crisp-dm-phases-and-tasks-ver2

[5] https://www.visualcapitalist.com/comparing-wealth-u-s-geographic-regions-time/

[6] https://www.business2community.com/brandviews/upwork/why-silicon-valley-techies-are-rushing-to-the-pacific-northwest-02076366